



University
of Glasgow

Mancy, Rebecca (2015) *Modelling persistence in spatially-explicit ecological and epidemiological systems*. PhD thesis.

<http://theses.gla.ac.uk/6219/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

MODELLING PERSISTENCE IN SPATIALLY-EXPLICIT ECOLOGICAL AND EPIDEMIOLOGICAL SYSTEMS

REBECCA MANCY

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE
COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

FEBRUARY 2015

© REBECCA MANCY

Abstract

In this thesis, we consider the problem of long-term persistence in ecological and epidemiological systems. This is important in conservation biology for protecting species at risk of extinction and in epidemiology for reducing disease prevalence and working towards elimination. Understanding how to predict and control persistence is critical for these aims. In Chapter 2, we discuss existing ways of characterising persistence and their relationship with the modelling paradigms employed in ecology and epidemiology. We note that data are often limited to information on the state of particular patches or populations and are modelled using a metapopulation approach. In Chapter 3, we define persistence in relation to a pre-specified time horizon in stochastic single-species and two-species competition models, comparing results between discrete and continuous time simulations. We find that discrete and continuous time simulations can result in different persistence predictions, especially in the case of inter-specific competition. The study also serves to illustrate the shortcomings of defining persistence in relation to a specific time horizon. A more mathematically rigorous interpretation of persistence in stochastic models can be found by considering the quasi-stationary distribution (QSD) and the associated measure of mean time to extinction from quasi-stationarity. In Chapter 4, we investigate the contribution of individual patches to extinction times and metapopulation size, and provide predictors of patch value that can be calculated easily from readily available data. In Chapter 5, we focus directly on the QSD of heterogeneous systems. Through simulation, we investigate possible compressions of the QSD that could be used when standard numerical approaches fail due to high system dimensionality, and provide guidance on appropriate compression choices for different purposes. In Chapter 6, we consider deterministic models and investigate the effect of introducing additional patch states on the persistence threshold. We suggest a possible model that might be appropriate for making predictions that extend to stochastic systems. By considering a family of models as limiting cases of a more general model, we demonstrate a novel approach for deriving quantities of interest for linked models that should help guide modelling decisions. Finally, in Chapter 7, we draw out implications for conservation biology and disease control, as well as for future work on biological persistence.

Acknowledgements

Not many people have the opportunity to study for a second PhD. I would like to thank the following, who have made it possible for me to do this all again.

Firstly, I want to thank my supervisors, Simon Rogers and Patrick Prosser, as well as my unofficial advisor, Dan Haydon. The level of trust that they have demonstrated in me, despite my sometimes hectic schedule, the competing demands on my time and my incessant temptations to take yet another new tack, have been really appreciated. I am also extremely grateful to my examiners, Adam Kleczkowski and Alice Miller, for their careful consideration of my work. The study would also not have been possible without funding from the Engineering and Physical Sciences Research Council, Grant Number EP/P505534/1.

A small group of people have shared the process of studying for a PhD with me. I would like to thank Anna Polychroniou, Hugues Salamin and team in room 322 for making sure I ate lunch on time, for their interest in my studies, and for listening to my rants and amusements! And more recently, Bilal Usmani, who has cheered me up on many occasions.

My friends and colleagues in Graham Kerr deserve a big thank you for keeping me sane and making sure that my work maintained some connection to real world biological systems!

I also want to thank my PhD students in the School of Education. I can't imagine that it is easy to know that your own supervisor is also studying for a PhD - albeit in a different discipline - and I want to show my appreciation for not making it all seem too weird!

Finally, I would like to express my deep gratitude to my partner, Konstantinos Angelopoulos, who in addition to practical help in all sorts of ways (including regular culinary delights!), supported me in applying to do a second PhD in the first place, and has put up with my excitement and exhaustion during this process, encouraging me throughout.

Thanks!

Declaration of Authenticity and Author's Rights

This thesis is the result of original research by the author, Rebecca Mancy. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

A black rectangular box containing a handwritten signature in cursive script, which appears to read 'Rebecca Mancy'.

Rebecca Mancy

Contents

1	Introduction	1
1.1	The importance of population persistence	1
1.2	The role of models in understanding persistence	3
1.3	Research aims	4
1.4	Thesis structure and contribution	5
2	Long-term persistence: concepts and models in ecology and epidemiology	8
2.1	Persistence in ecology and epidemiology	8
2.1.1	Why persistence matters	8
2.1.2	Models and reality	9
2.2	Modelling decisions	11
2.2.1	Simple example model	12
2.2.2	Deterministic or stochastic?	13
2.2.3	How to model time	14
2.2.4	How to model the population	15
2.2.5	How to model states	17
2.2.6	Relationships between modelling decisions	17
2.2.7	Modelling decisions as a computational biology question	18
2.3	Persistence	22
2.3.1	Operational definitions in deterministic models	24
2.3.2	Operational definitions in stochastic models	25
2.3.3	Operational definitions in data-driven work	26
2.3.4	Other related definitions	26

2.3.5	Persistence and model types	28
2.4	Probability theory and simulation concepts	29
2.4.1	Gillespie algorithm	29
2.4.2	The quasi-stationary distribution	30
2.5	Models and persistence definitions in this thesis	33
2.6	Conclusion	36
3	Discrete and continuous time simulations of spatial ecological processes predict different final population sizes and interspecific competition outcomes	37
3.1	Introduction	38
3.2	Modelling decisions of cellular automata	39
3.3	Experimental protocol	41
3.3.1	Continuous time model	41
3.3.2	Cellular automaton update schemes	42
3.3.3	Rate conversion	42
3.3.4	General model parameters	43
3.4	Experiment 1: Single population	43
3.5	Experiment 2: Interspecific competition	48
3.6	Discussion and conclusions	53
4	The contribution of a patch to persistence in a stochastic metapopulation model	56
4.1	Motivation	57
4.2	Characterising patch value	58
4.3	Materials and methods	60
4.3.1	Model	60
4.3.2	Experimental setup	62
4.3.3	Persistence measures	65
4.3.4	Predictors of patch value	66
4.4	Findings	67
4.4.1	Predictors of time to extinction	67
4.4.2	Refined estimation of η	75

4.4.3	Predictors of metapopulation size	78
4.4.4	Predicting sensitivity to patch removal	82
4.5	Discussion	83
4.6	Concluding remarks	85
5	Simulating quasi-stationary behaviour for heterogeneous systems	86
5.1	Introduction	87
5.2	Materials and methods	93
5.2.1	Model	94
5.2.2	Simulation algorithm	95
5.2.3	Compressions of the QSD	97
5.2.4	Experimental setup	104
5.3	Accuracy of the full simulation	106
5.3.1	Algorithm refinement for least persistent systems	106
5.3.2	Trend as n increases	107
5.4	Accuracy of QSD compressions	109
5.4.1	Relationship between compressions	110
5.5	Selection of a compression	114
5.5.1	Compression for overall accuracy	114
5.5.2	Compression for extinction times and extinction paths	117
5.5.3	Compression for occupancy patterns	120
5.6	Discussion	120
5.6.1	Concluding remarks	125
6	Persistence thresholds in related multi-state ecological and epidemiological models	126
6.1	Introduction	127
6.2	Compartmental models	128
6.3	A family of compartmental models	131
6.3.1	Logistic growth and Levins models	131
6.3.2	Spatially realistic Levins model	132

6.4	Links between different multi-state models	135
6.5	Persistence thresholds and their problems	137
6.5.1	Allee effects	139
6.6	Derivation of SIIS thresholds	141
6.6.1	Levins model	141
6.6.2	Levins model with strong Allee effect	143
6.6.3	Levins model with weak Allee effect	145
6.6.4	SRLM SIIS	146
6.6.5	Comparison of thresholds	148
6.7	Discussion and conclusions	149
7	Discussion	154
7.1	Overview	154
7.2	Implications for future work	160
7.2.1	Model types and persistence	160
7.2.2	Questions arising from possible applications	162
7.3	Concluding remarks	163
A	Chapter 1	164
A.1	Great principles of computing science	164
B	Chapter 3	169
B.1	Algorithms	169
B.1.1	<i>Gill</i>	169
B.1.2	RF_d2S algorithm	173
B.1.3	$RR1S$ algorithm	174
B.1.4	Multiple births RF_d2M and $RR1M$	175
C	Chapter 4	177
C.1	Alternative predictors of ∇T_m	177
C.1.1	p_i^π	177
C.1.2	p_i^*	177

C.1.3	A_i	178
C.1.4	V_i	179
C.2	Alternative predictors of ∇S^π	179
C.2.1	p_i^π and proportion of S^π due to patch i	179
C.2.2	V_i and V_i^π	180
C.2.3	p^*	180
D	Chapter 5	183
D.1	QSD comparison statistics	183
D.1.1	Pearson correlation coefficient	184
D.1.2	Root mean square error	184
D.1.3	Angle between vectors	185
D.1.4	Kullback-Leibler divergence	185
D.1.5	Findings	186
D.2	Robustness testing	186
E	Chapter 6	189
E.1	Correspondence between Levins and logistic growth model	189
E.2	Derivation of non-standard form of <i>SIIS</i> Levins model	190
	Bibliography	191

List of Tables

2.1	Overview of questions and definitions of persistence used in studies described in the thesis.	35
3.1	Maximum time step size giving indistinguishable final population sizes as <i>Gill</i> (2-tailed student t-test, unequal variances at the 5% level); – indicates that final population sizes differed for all of the time steps considered; N/A indicates populations that went extinct for both simulation approaches; * indicates that this algorithm did not correctly predict extinction for these values.	47
4.1	Symbols used in the SRLM	60
4.2	Parameter values	64
5.1	Symbols used in the SRLM	94
6.1	Illustrations of different compartmental models with, in the figures, epidemiological terminology (above) and ecological terminology (below).	130
6.2	Persistence thresholds for the family of compartmental models. The strong Allee effect has two versions: (a) where the threshold is on $p_{I_1} + p_{I_2}$ and (b) where is it on p_{I_1} only.	148
A.1	Application of principles of design	165
A.2	Application of principles of mechanics	166
A.3	Application of practices	167
B.1	Symbols used in algorithms	170

List of Figures

2.1	Illustration of the relationship between model types resulting from decisions about how to aggregate over individuals and space.	16
2.2	Compartmental model of population disease states.	23
2.3	Conceptual illustration of three scenarios following introduction of an infectious agent. Redrawn from Gubbins et al. (2000); Gilligan and van den Bosch (2008).	24
2.4	Visual representation of models used in the thesis	34
3.1	Box plot showing final population sizes for death-to-birth ratios δ under <i>Gill</i>	44
3.2	Time series plot showing mean and range for $\delta = 0.7$ (lower line) and $\delta = 0.6$ (upper line) under <i>Gill</i>	44
3.3	Plots of (a) difference in average population size (including extinct runs) for <i>Gill</i> – <i>RF_d2S</i> (shading highlights difference in final population size) and (b) percentage algorithm bias (upper value in each cell; shading highlights size of algorithm bias) and number of extinct runs out of 100 under <i>RF_d2S</i> (lower value). Negative values represent underestimates compared with <i>Gill</i>	46
3.4	Percentage algorithm bias for <i>RF_d2M</i> (upper value in each cell) and number of runs extinct out of 100 (lower value). Negative values represent underestimates compared with <i>Gill</i> ; shading highlights absolute algorithm bias.	47
3.5	Box plot showing final population sizes for the two species under different death-to-birth ratios δ under <i>Gill</i>	50
3.6	Interspecific competition for <i>RF_d2S</i> (top panel) and <i>RF_d2M</i> (bottom panel). The upper two values in each cell represent the final population size for species one and two respectively with bold type used to highlight the larger of the two values. The bottom value in each cell gives percentage of runs in which species one dominated at time 1000. Shading indicates parameter sets giving the same qualitative predictions as <i>Gill</i>	51
3.7	Time series plots (mean and range) for <i>Gill</i> , <i>RF_d2S</i> and <i>RF_d2M</i> , $\delta = 0.1$	52

4.1	Overall experimental procedure.	63
4.2	The dependence of T_m on persistence as determined by changes in values of e (note logarithmic scale; each line represents a landscape).	68
4.3	The relationship between $(p_i^\pi)^2$ and ∇T_m for $n = 10$. Parameters $\alpha = 1$ and $c = 1$ fixed; e is selected to change overall system persistence in relation to λ_M	69
4.4	Estimator bias of $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$	70
4.5	Root mean squared error of $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$	70
4.6	Pearson correlations for $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$	71
4.7	Spearman correlations for $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$	71
4.8	The relationship between ∇T_m and $(p_i^\pi)^2$ for different values of $\zeta_{im}, \zeta_{em}, \zeta_{ex}$ for log normally distributed patch areas. The top figure shows high persistence, the middle figure intermediate persistence $\lambda_M = \delta$ and the bottom figure shows low persistence; $n = 5$ and $\alpha = 1.0$. The coloured boxes in the top left of each scatter plot show root mean squared error (RMSE) (see top box; light shading indicates low RMSE) and Spearman correlation (bottom box; dark shading indicates high correlations). Facets labelled 0,1,2 on the right-hand side of plots give ζ_{em} values.	73
4.9	The relationship between ∇T_m and V_i for different values of $\zeta_{im}, \zeta_{em}, \zeta_{ex}$ for log normally distributed patch areas and $n = 5$ and $\alpha = 0.5$. Explanation as for Figure 4.8	74
4.10	Histograms of the RMSE in predicted η values for the holdout set using (a) the baseline prediction of $\eta = 2$ and (panels b,c,d) multiple regression models fitted to the training set. The three models are fitted to the training set using a multiple linear regression of η on the metapopulation parameters, where η is found as the slope of the simple regression line of $\log(\nabla T_m)$ on $\log(p_i^\pi)$	76
4.11	The proportion of the total area occupied in the QSD as a function of the persistence parameters.	78
4.12	The relationship between potential measures and ∇S^π	79
4.13	The relationship between U_i and ∇S^π	80
4.14	Estimator bias of U_i and V_i as predictors of ∇S^π	80
4.15	Root mean squared error of U_i and V_i as predictors of ∇S^π	81
4.16	Pearson correlation between U_i and V_i as predictors of ∇S^π	81

4.17	Spearman correlation between U_i and V_i as predictors of ∇S^π	82
4.18	The effect of patch removal on T_m for different initial mean occupancy levels in the QSD ($n = 10$). Each line represents the removal of patch zero for one landscape.	82
4.19	The effect of patch removal on mean metapopulation size for different initial mean occupancy levels in the QSD ($n = 10$). Each line represents the removal of patch zero for one landscape.	83
5.1	Overall experimental protocol.	94
5.2	Illustration of system storage for a system with five patches after 10 census points for full storage and the three compressions, given states labelled such that system state one is [10000], for the hypothetical chain: [11000] [01000] [11000] [11000] [10000] [11000] [01000] (then extinction and uniform random jump to full state) [11111] [11111] [11111]. In independent patches by cluster, three clusters are used and the first records states [01000] and [11000]. Colours for the different forms of storage are used throughout the chapter. .	100
5.3	Sample landscape with five patches and log normally distributed patch areas.	106
5.4	Two runs of the baseline algorithm for the landscape shown in Figure 5.3 and parameters giving low persistence.	107
5.5	Left panel: State probabilities of exact QSD and full simulation (refined algorithm). Right panel: Scatter of state probabilities of full simulation against exact QSD. Both for the landscape shown in Figure 5.3 ($e = 0.1$, $\alpha = 0.5$). .	108
5.6	Algorithm accuracy in the form of median Pearson correlation between the true QSD and full simulation for different values of n and persistence e , shown on the right-hand axis (in all figures, persistence decreases as we move down the panels).	109
5.7	Algorithm accuracy of the full simulation, independent patches and independent patches by occupancy.	111
5.8	Comparison of compressions with similar memory footprint.	112
5.9	Comparison of adjustments to maximum number of clusters.	113
5.10	Proportion of available clusters used for landscapes with exponentially distributed patch areas.	114
5.11	More detailed comparison of independence by occupancy with clustering with $K_{max} = n^2$	115
5.12	Median total error calculated as square root of sum of square errors.	117

5.13	Proportional absolute error in mean time to extinction from quasi-stationarity.	118
5.14	Mean of absolute error in mean time to extinction from quasi-stationarity, as a proportion of true T_m .	119
5.15	Probability occupancy (darker shading representing higher probabilities) of each patch in the two most common clusters in the cluster representation of the QSD for an example landscape with 5 patches and $e = 0.1$, and counts of the number of census points (out of 452537 census points during the simulation, using n clusters).	121
5.16	Probability occupancy of each patch (darker shading representing higher probabilities) in the two most common clusters in the cluster representation of the QSD for an example landscape with 5 patches and $e = 0.1$ with counts of the number of census points, using n^3 clusters.	122
5.17	Median relative difference in estimates of T_m from independent patches by occupancy compared with those from the exact QSD (normalised by exact T_m , for log normally distributed patch areas) and showing 95% of the values (darker ribbon) and the full range (lighter ribbon).	123
6.1	Relationship between compartmental models, in which c_{I_1}, c_{I_2} give the colonisation rates from I_1 and I_2 respectively; similarly for the recovery rates from these compartments r_{I_1}, r_{I_2} .	137
6.2	SIIS model. Dynamics and rates for a single patch i .	146
C.1	The relationship between p_i^π and ∇T_m (parameters as in Figure 4.3).	177
C.2	The relationship between p_i^* and ∇T_m (parameters as in Figure 4.3).	178
C.3	The relationship between A_i and ∇T_m (parameters as in Figure 4.3).	179
C.4	The relationship between V_i and ∇T_m (parameters as in Figure 4.3).	180
C.5	The relationship between potential measures using the raw p_i^π values and S^π .	181
C.6	The relationship between \tilde{V}_i and S^π .	182
C.7	The relationship between p^* and S^π .	182
D.1	Box plots showing the statistics considered for vectors of length 2^n .	184
D.2	Comparison of adjustments to clustering parameters with $K_{max} \approx n$.	187
D.3	Algorithm accuracy for clustering algorithm with n^3 and different values of α .	188

Chapter 1

Introduction

1.1 The importance of population persistence

Understanding population persistence is important for both the basic science of population biology and for management purposes. From an ecological perspective, we typically wish to be able to predict species persistence and manage resources to minimise extinction risk; in a disease ecology or invasive species context, we wish to be able to anticipate pathogen persistence and design interventions to work towards reductions in prevalence or elimination.

An iconic theme in the literature is that of understanding the effects of climate change on species persistence. For example, the Nature group launched a new journal, *Nature Climate Change*, in 2007, specifically dedicated to understanding climate change and its consequences, noting that this represents ‘a scientific challenge of enormous importance to society’. Climate change is expected to lead to a heightened risk of species extinction risk (e.g. Thomas et al., 2004). However, climate change interacts with other anthropogenic drivers, especially in the form of land-use change or habitat loss and fragmentation and competition with invasive species, themselves possibly driven by climate change (Thomas et al., 2004). Further threats may be seen in the emergence of novel infectious diseases and the collapse of populations of pollinator species that are expected to lead to long-term declines in the species that depend upon them (Gonzalez, 2013). The interactions between these factors may well be complex, counterintuitive or highly contextual (e.g. Lafferty, 2009).

Commitment to the investigation of environmental change and its effects is also reflected in research funding. For example, it can be seen in the decision to make ‘Living with Environmental Change’ (LWEC) a priority theme for all of the UK research councils, and the formation of the LWEC partnership of 22 major UK public sector funders and users of environmental research, which includes the research councils and central government departments, and aims to ensure that key decision-makers have access to the environmental

research required to ‘mitigate, adapt and benefit from environmental change’.

The impacts of environmental change on species persistence also have important societal implications (Gonzalez, 2013). Prominent examples include the impacts of climate change on crop harvests, especially in interactions with declines in pollinator populations. However, the notion of *ecosystem services* is increasingly recognised as important, emphasising the role of ecosystems and biodiversity in contributing to human health and wellbeing (Daily et al., 2009). The potential for environmental and societal change to lead to the emergence of new infectious diseases, or extension in the endemic range of existing diseases, is also of concern for humans (Cohen, 2000; Patz et al., 2005) and wildlife (Daszak et al., 2001).

In contrast to the increasing habitat fragmentation scenario currently being played out in relation to non-human, non-livestock species in the ecological context, examining host populations from an epidemiological perspective finds that these are typically becoming more strongly connected, although connections typically remain heterogeneous. This is especially true of human populations, but is also applicable to livestock that may be transported greater distances because of improved travel connections and a more internationalised market. If habitat fragmentation is typically thought to lead to species declines, increased host population connectivity speeds up disease transmission, typically increasing the size of epidemics (e.g. Balcan et al., 2009). These effects are likely to support emergence of novel diseases and may also increase persistence (Wilson, 2004) if spread is not so fast that it leads to burnout of susceptible individuals. The impacts of this pattern of increased human connectivity in conjunction with reduced connectivity of wildlife habitat patches, present serious concerns for the scientific and policy-making communities.

Designing interventions to optimise species persistence characteristics in the face of different levels of connectivity - for either increased persistence or elimination - requires an understanding of how best to direct resources towards the areas of habitat or host subpopulations where their impact will be greatest (Viana et al., 2014). In the conservation biology literature, site selection algorithms are discussed as a way to choose between possible habitat areas for reserve networks (Cabeza and Moilanen, 2001). In the epidemiological literature, work to develop effective schemes for the allocation of vaccines is also underway (Beyer et al., 2012). Nonetheless, despite the importance of these problems, speaking in the context of conservation Gonzalez (2013, p206) points out that it often remains ‘difficult to predict when and where species losses will occur, and this hampers efforts to prevent them’. This thesis contributes to making inroads into the problems of understanding persistence in heterogeneous landscapes and populations.

1.2 The role of models in understanding persistence

There are fairly obvious practical problems involved in investigating population persistence from a classical experimental or even observational perspective. Taken at a species level, either the species continues to exist, or it goes extinct. At best, we may be able to calculate the time to extinction for this species when extinction occurs or track population sizes prior to this. As Harnik et al. (2012) puts it, the impact of rarity and population size on population extinction ‘cannot be determined through observations of species that are not yet extinct’ (p4969). This means that extinction is typically difficult to predict. An obvious solution to this problem is to consider the problem at smaller scales by disaggregating the population into subpopulations and tracking their persistence and extinction, such that a larger number of data points can be collected. Nonetheless, for many species, timescale remains a serious problem as the waiting time until subpopulation extinction still typically exceeds that of grants, and often that of record-keeping more generally. Furthermore, although disaggregation into smaller and smaller (sub)populations reduces the persistence time of each component population and thus facilitates the collection of reasonably-sized datasets, doing so raises the question of how to re-aggregate data in order to predict persistence times of the metapopulation.

Modelling provides one way to gain some traction on this question. A large number of populations can be simulated much more rapidly than their real-world counterparts, generating large datasets from which patterns in the relationship between population and metapopulation persistence can be inferred. Because it is relatively cost-effective to generate such (simulated) datasets, the effects of changing model assumptions can be investigated, allowing us to test the generality of findings. Although it remains a challenge to map these findings to real world systems, modelling may also suggest the kinds of data that would be most informative to collect, not just for predicting extinction *per se* but also for understanding correspondence with particular patterns identified in the modelling context.

In reality, we have more measurement options at our disposal than simply recording the presence or absence of an organism or the date of its extinction. We might, for example, choose to measure the area inhabited by an organism or the population density; in a fragmented landscape, we could measure the proportion of time that an area is occupied. These possibilities raise the question of which of the potential measures provides the most valuable information about population persistence. Here again, modelling and simulation approaches can be very valuable by allowing us to investigate the relationship between measurable characteristics of a population and the information that they can provide about persistence. Answers to this question have the potential to have major contributions to the most efficient use of resources employed in data collection. This is particularly valuable because surveillance and monitoring are expensive (McDonald-Madden et al., 2010) and these activities may lose out to

interventions (Stankey et al., 2005; Viana et al., 2014). By being clearer about what kind of data collection is required, we can better direct limited resources.

Another problem, that is situated between theory and practice, concerns the way in which persistence is defined. When authors define systems as persistent, they typically use terms that imply a certain permanence. For example, in defining disease reservoirs, Haydon et al. (2002, p1469) referred to a system in which the pathogen can be ‘permanently maintained’, while Ashford (2003, p1495) described it as one in which an infectious agent ‘survives indefinitely’. However, we still need to translate these theoretical distinctions into a form that remains at once faithful to the theoretical definitions and operational as a way of guiding data collection. For example, we may decide to use an operational definition based on the time to extinction or the probability of extinction within a target period, from a pre-specified starting condition (Ludwig, 1999; Viana et al., 2014). As seen in the examples provided by Odenbaugh (2005), the formalisation of concepts that is required by the modelling process provides us with this opportunity to sharpen our operational definitions of persistence, as well as the relationship between them and their link with theoretical definitions.

As a result, modelling can help us to understand persistence by simultaneously reducing our reliance on the collection of large real-world datasets, by guiding our real-world data collection towards the most productive measures, and assisting us in the formalisation of useful operational definitions of persistence.

1.3 Research aims

The broader aims guiding the research reported in this thesis are to develop our understanding of biological persistence, and to explore the role that modelling can play in this development. This development is conducted through four studies that explore the effects of (1) modelling decisions that relate to the relationship between the processes and structures of the system and those embodied in the model and (2) measurement decisions about how to capture and characterise persistence. The two themes of persistence and modelling decisions thus serve as leitmotifs, reappearing in the different sections of the work. The main modelling decisions compared are those of whether time is modelled discretely or continuously, different ways of characterising the contribution of a patch to metapopulation persistence, different ways of representing the distribution of system states in the long run, and the effect of modelling differing numbers of states for each patch in a metapopulation. Persistence is characterised in different chapters as continued existence until a pre-defined target time, mean time to extinction, patch occupancy probability distributions in the long run, and as the existence of a stable non-trivial equilibrium in a deterministic model. The particular research questions addressed in each chapter are identified in the context of the particular piece of work, and

the contribution of the answers to these questions to the overall aims is discussed in the main discussion in Chapter 7.

1.4 Thesis structure and contribution

The main body of this thesis is structured as follows. The thesis is organised around the central theme of modelling decisions and their effect on the conclusions that can be drawn from studies relating to persistence. The main body of the thesis consists of a series of semi-independent chapters, written in extended paper form, each with its own literature review that is used to identify and justify specific research questions, a description of the methods employed in answering them, a presentation of the findings and a discussion of their implications. As a result, some areas of theoretical background are repeated and the reader can progress rapidly through these where indicated.

Chapter 2 provides background theory and ideas in order to assist the reader in understanding the main contributions of the thesis. Because the work is interdisciplinary, at the juncture of computing science, mathematical biology, and ecology, an attempt has been made to provide sufficient background for readers with a training in any of these areas; as a consequence, many readers will find that other aspects are covered in more depth than is required for their own purposes. The chapter has two main sections. In the first section, some key concepts for modelling biological populations are introduced alongside a proposed scheme for model categorisation based on the decisions to model elements as continuous or discrete and the level of aggregation included. There is also a subsection on the relationship between the study of modelling decisions and research in computing science, and its fit with the area of computational biology. The second section focuses on current definitions of persistence from the biological literature of relevance in both deterministic and stochastic contexts. These are used to argue that greater consistency of terminology would be beneficial for research in this area. We make the case for research on the theme of population persistence and the importance of considering the relationship between modelling decisions and their effects on conclusions about persistence.

Chapter 3 consists of a stochastic simulation study using a cellular automaton approach, and explores the effects on persistence of a single species and in the context of interspecific competition, of the decision to model time as continuous or discrete. We use a simple notion of persistence, sampling at a pre-specified census point. We show that there are quantitative and qualitative differences in persistence between populations simulated in discrete and continuous time and argue that this modelling decision is important. We also show that the use of a pre-specified time horizon as a measure of persistence can be misleading. This chapter has previously been published as Mancy et al. (2013) and provides a novel contribution to

the ecology literature by highlighting the importance of the decision to model processes in discrete or continuous time, and specifically the difficulties of approximating continuous spatial processes by discrete time simulations, especially in systems with more than one species. From a computational biology perspective, it provides a comparison between the accuracy of existing algorithms for simulating birth-death processes on a regular lattice, while also employing a new version of the continuous time algorithm with reduced space complexity compared with the classical approach employed for this kind of system.

Chapter 4 focuses on the question of how to model the contribution of a patch to metapopulation persistence. We use a stochastic version of a metapopulation model with patch heterogeneity with a finite number of patches that become colonised and go extinct dynamically. As our measure of persistence, we employ the mean time to extinction from quasi-stationarity. The quasi-stationary distribution (QSD) is a description of the distribution of system configurations, conditioned on non-extinction, and thus captures the expected long-run states of the system. We find expressions for the contribution of a patch to mean time to extinction from quasi-stationarity, as well as population size. The chapter contributes to the theoretical biology literature by providing an expression for the contribution of a patch and raising further questions about the exact relationship between patch occupancy probability and patch value. From a computational science perspective, it provides an illustration of the use of a computational study to uncover a regularity suggesting an intervention design heuristic which could be employed without recourse to further complex computational work. This work also has the potential for significant contributions to real-world applications since it suggests a relatively robust estimate of the contribution of a patch or population based on only readily available data.

Chapter 5 focuses on the question of how to simulate the quasi-stationary distribution, as the kind of study conducted in Chapter 4 is currently impossible for systems with a large number of states. In this chapter, we investigate possible approaches for storing a compressed version of the QSD in order to provide an accurate representation of the QSD without requiring full state storage. We test different storage schemes based on different patch independence assumptions. We find that none of these is entirely robust to increases in the number of patches, but for specific uses, it is possible to select an appropriate scheme. The work has applications to metapopulations that are larger than those for which numerical solutions are feasible. From a computational biology perspective, the main novel contributions of this chapter are in the findings about the accuracy of the schemes tested, and in the development of the algorithms for simulating the QSD and its compressions. The findings provide a practical tool to guide the modelling decisions of researchers interested in the QSD for these kinds of systems, while raising engineering questions related to the possible ways to improve upon the compression algorithms or models considered.

In Chapter 6, we consider a range of deterministic metapopulation models as a way to in-

investigate the extent to which these can provide useful information about the persistence of real-world systems. We introduce a framework for explicitly linking related models, deriving the persistence threshold for the most general models and allowing parameter values to tend to appropriate limits to recover more specific models. The main novel contribution of this chapter is theoretical and rests on the framework for working with models that can be considered as limiting cases of a model that subsumes them, and that underpins the derivation of the persistence thresholds. This approach could be applied more broadly, especially to guide modelling decisions about the states to include in compartmental models. A number of the persistence thresholds are also new, and those including an Allee effect raise possibilities of using deterministic models to investigate persistence. The work in this chapter arises from considerations developed in connection with our contributions to the development of the notion of reservoir capacity in Viana et al. (2014), co-authored during the period of the PhD. It raises new questions relating to the connection between persistence thresholds in deterministic models and their stochastic counterparts, that could be explored computationally.

In Chapter 7, the focus is on a discussion of the findings from the main studies reported in this thesis in relation to the overarching aims to understand persistence and the role of modelling in this endeavour. The chapter begins with an overview of the findings in relation to their possible use to guide decisions about how to model and to characterise persistence, and their contributions to computational biology. This is followed by a discussion of possible applications of the work in conservation biology and infectious disease control, and implications for future work on persistence. This section also includes a discussion of caveats and suggestions for future research.

Chapter 2

Long-term persistence: concepts and models in ecology and epidemiology

In this chapter, background literature is discussed in relation to two frameworks: the first focuses on modelling decisions and model types in ecology and epidemiology; the second on meanings of persistence. Each of these has been developed to serve as an organisational schema for discussion of these ideas. We begin with a section on the importance of persistence in ecology and epidemiology, and move on to present the two frameworks. In the closing sections of the chapter we present two key probability theory concepts - the Gillespie algorithm and the quasi-stationary distribution - and provide an overview of our application of the ideas described here to the studies presented in the thesis.

2.1 Persistence in ecology and epidemiology

2.1.1 Why persistence matters

The related notions of persistence and extinction are of fundamental importance in population ecology and infectious disease epidemiology. In population ecology, we are often interested in whether a species or community will persist in the short or long term, and how the answers to these questions may vary as a result of changing ecological conditions such as habitat deterioration, climate change or conservation actions. Such questions are therefore important for effective resource management in a range of biological systems. The response of species to habitat disturbance such as habitat fragmentation and climate change constitute important contemporary questions in conservation biology (see e.g. Thomas et al., 2004).

In infectious disease epidemiology, the control, elimination and ultimate eradication of infectious disease constitute the main aims of interventions. Although good progress has been made towards elimination of formerly endemic diseases, the final push towards worldwide eradication of many diseases remains elusive, of which polio and rabies constitute iconic examples (Pallansch and Sandhu, 2006; Lankester et al., 2014). In addition, other diseases such as dengue are rapidly gaining new territory to become endemic in a range of regions across the globe (Bhatt et al., 2013). In the case where interventions to prevent emergence (or re-invasion) fail, understanding persistence becomes important.

Nonetheless, although persistence and extinction are issues of fundamental importance in population ecology and infectious disease epidemiology, their dynamics remain relatively poorly theorised, at least in comparison with invasion dynamics. Persistence has usually attracted attention in the form of considerations of time to extinction and extinction rates. However, relevant theoretical definitions and their operational counterparts remain inconsistent in the literature, and there is little consensus on terminology.

Taken together, the empirical work presented in this thesis advances research on persistence directly and through an examination of the implications of specific modelling decisions for the investigation of persistence. In this chapter, we provide background information and make the case for research on persistence in ecological and epidemiological settings. The first section of this chapter examines the relationship between models and the ‘real’ world, focusing on modelling decisions of relevance to the investigation of persistence in population and disease ecology. We then discuss existing notions of persistence and endemicity in the literature, and their associations with different modelling paradigms. We argue that greater coherence in the use of terminology relating to persistence would be valuable and propose a model and associated terminology for work in this area.

2.1.2 Models and reality

Models of biological systems necessarily involve simplifying or idealising assumptions, characterised by Takacs and Ruse (2011, p21) as ‘often patently false or only partially true’. Nonetheless, Odenbaugh (2005, p251) argues that ‘models that are inaccurate can still lead to significant truths,’ arguing that the same model might be used for a variety of purposes and must be evaluated against the uses to which it is put, concluding that ‘insofar as models accomplish those functions, we can count those models as successful.’ Determining which assumptions are likely to be most productive typically requires evaluating tradeoffs that depend on the aims of the model and how it is analysed or used to answer biological questions. As Pincock (2012) points out, the mathematics itself is unable to tell us when these simplifications are appropriate for particular modelling applications. Biologists often ask themselves questions such as the appropriate level of aggregation for modelling a system, or the number

of biological levels that should be included (Orzack, 2012). For example, one may consider whether it is necessary (or even desirable) to model individuals explicitly when investigating ecological stability. Deciding how best to simplify or idealise a system when constructing a model is therefore an important question in scientific modelling.

Although modelling is typically associated with generating explanations and predictions, Odenbaugh (2005, p236) argues that models have at least five basic purposes in theoretical ecology:

1. Models are used to *explore possibilities*.
2. Models give scientists simplified means by which they can *investigate more complex systems*.
3. Models provide scientists with *conceptual frameworks*.
4. Models can be used to *produce accurate predictions*.
5. Models can be used to *generate explanations*.

According to this framework, when models are used to *explore possibilities*, their use revolves around exploring hypotheses relating to what a biological system might do in particular circumstances. This might result from changing parameter values to evaluate the effect¹. When they are used to *investigate more complex systems*, very simple and deliberately unrealistic models are employed as ‘baselines’ and investigations focus either on evaluating which assumptions are problematic by comparing with real-world data, or on understanding the implications of adding or changing the assumptions in order to better reflect real-world phenomena. In other circumstances, they may be used to provide *conceptual frameworks* that can be used to ask questions and conduct experimental investigations. This usage is sometimes thought of as a side-effect of the process of model development, but could be deliberate: it refers to the development of concepts that result from the formalisation of ideas required for model building. For example, it is in this way that models contributed to the development of the concepts such as *connectance* and *species-deletion stability*. As Caswell (1988, p35) points out, the importance of these, and related concepts, to complexity and ecosystem stability only became clear as the original verbal theory was developed into mathematical models. In many circumstances, models are used for the purpose of *producing accurate predictions* in the sense that the models match state variables that can be measured in the real world. Finally, models can be used to *generate explanations* in the sense of providing possible mechanisms by which a phenomenon occurs. Although the distinction between

¹The distinction between this category and ‘generate explanations’ seems unclear insofar as these explorations are characterised as leading to ‘how possibly’ explanations.

some of these categories remains rather unclear, the point that models can serve multiple purposes is well made.

In summary, models necessarily embody simplifications of real-world systems, and the modelling decisions about which simplifications to make are important; however, these decisions can only be made in relation to the aims of the model. That is, as Richard Levins (1985, p8, cited in Odenbaugh, 2005) writes,

Modelers always must keep in mind that the utility of their construct depends on the particular purpose for which it was built. There is no such thing as *the true model* of a system but only more or less adequate representations of the system.

2.2 Modelling decisions

Models of population persistence need to represent both time and the population in some way. However, each of these aspects can be modelled in a range of ways and decisions about how they are modelled can have important implications for the conclusions that can be drawn from the models. An additional modelling decision that is particularly important for studying population persistence is whether to use a deterministic or stochastic model.

Several authors have discussed modelling paradigms and approaches used in ecological and epidemiological modelling. A majority of these presentations progress by describing prototypical examples of each paradigm, and then comparing and contrasting these (see Keeling and Rohani, 2007; Barrat et al., 2008; Bolker, 1997; Lion and van Baalen, 2008; Jorgensen, 2008). For example, in relation to epidemiological modelling, Barrat et al. (2008) distinguish between homogeneous mixing models, models that incorporate social structure, contact network models, multiscale models (patch or meta-population) and agent-based models, describing these as models at different ‘scales’ (p181). Focusing on spatially-explicit epidemiological models, Keeling and Rohani (2007, Chapter 7), distinguish between metapopulation models, lattice-based models, continuous-space continuous-population models, individual-based models and networks. Bolker (1997) focuses on ecological models and distinguishes between continuous-density models, lattice and patch models (spatially-structured models) and simulation models, and introduces moment equation approximations. The above authors organise their discussion around paradigmatic examples and although there is some overlap between the types of models described by these authors, there is also an apparent arbitrariness in the classification schemes and examples discussed. Some authors have attempted to classify model types according to their features (e.g. Gertsev and Gertseva, 2004). Finally, Berec (2002) begins by establishing a framework that is used to compare model types based on modelling decisions.

We take a framework-based approach that is similar to that of Berec (2002) as this creates an explicit structure that can be used to inform modelling decisions, and that explicitly focuses on the elements of the biological system to be modelled. In contrast to Berec (2002), who considers continuous and discrete approaches to modelling time and space, we first address the decision about how to model time (as discrete or continuous), and then take a hierarchical approach to considering how the population and spatial aspects are modelled in respect of the level of aggregation employed. In a separate section, we explore the decision of whether to model population persistence using deterministic or stochastic processes.

2.2.1 Simple example model

Traditionally, researchers in population ecology have used differential equations to model population dynamics, changes in the size and composition of populations and the processes that cause these changes. For example, one of the most important models in population biology is the Verhulst-Pearl logistic growth model. This model captures the dynamics of population size in an environment where resources are limited. When resources are plentiful, individuals give birth at a rate b , and die at rate d . Resource limitations are modelled in the form of a *carrying capacity*, usually denoted K , representing the maximum population that the habitat can support in the long term. Near the carrying capacity, the birth rate falls and the death rate increases. Rates b and d are the intrinsic instantaneous birth and death rates *per capita*, and are thus multiplied by the number of organisms in the population to give the overall population rates. Defining $r = b - d$ as the maximum rate of growth (*per capita*), and x as the number of organisms in the population, the rate of change of the number of organisms in the system is usually expressed as

$$\frac{dx}{dt} = rx \left(1 - \frac{x}{K}\right). \quad (2.1)$$

Dividing both sides of the equation by K and defining $P = x/K$, we can convert the equation describing the rate of change of the *number* of organisms into one capturing the dynamics of population *density* (i.e. proportion of the carrying capacity that is exhausted; $1 - P$ is therefore the proportion of the resources still available). We can then write

$$\frac{d}{dt} \frac{x}{K} = r \frac{x}{K} \left(1 - \frac{x}{K}\right) \iff \frac{dP}{dt} = rP(1 - P) = (b - d)P(1 - P). \quad (2.2)$$

Because the equation describes the rate of change of population density, solving it for P gives the time evolution of the population density P . This model is mathematically equivalent to the Levins metapopulation model, and to the standard susceptible-infectious-susceptible model in epidemiology, as explained in more detail in Chapter 6.

This equation demonstrates the main features of the traditional approach to dynamic population modelling. Firstly, the model is deterministic. In it, both time and the population are modelled as continuous variables (even in Equation 2.1 because nothing precludes the possibility of non-integer organism counts). The population is modelled simply as the number of organisms: organisms are neither discrete nor heterogeneous. Finally, if we think of K as a measure of the number of available sites, then x is the number of sites that are occupied and $K - x$ is the number of unoccupied sites, and each site has two states: occupied or vacant. We now consider alternative modelling decisions that could be made.

2.2.2 Deterministic or stochastic?

A key modelling decision, and one that takes on a particular importance for population persistence, is that of whether to model the processes as deterministic or stochastic. Although the issue of whether the universe is deterministic or stochastic constitutes an unresolved problem in philosophy of science (i.e. the question of whether ontological randomness exists), the limits of our knowledge and the scientific process of knowledge generation mean that we often wish to include stochasticity in our models (i.e. we model epistemological randomness). Various ways to partition stochasticity exist in the literature. For example, Bolker (2008) distinguishes between three sources of random variability: measurement error, demographic stochasticity and environmental stochasticity, of which the last two concern variability or unpredictability in the underlying process itself. Measurement error refers to variability due to difficulties of accurate measurement of real-world systems. Demographic stochasticity refers to differences in the endogenous processes of the ecological system, such as variability in the numbers of births and deaths in any instantiation of a controlled experiment (i.e. variability that is not related to the variable that is controlled). Demographic stochasticity is endogenous to the system in that individuals in the system have some control over it (e.g. in a predator-prey model, the predator has control over prey death). Environmental stochasticity refers to unpredictabilities in exogenous processes such as climate that occur independently of the system that we are modelling, and that affect ecological dynamics.

In large-scale systems, the effects of demographic stochasticity may not be very important as random effects may be expected to cancel. As a result, deterministic models capture the dynamics of large populations better than small populations. However, in smaller populations, random effects are more important, such as when random fluctuations in population size lead to extinction.

Stochasticity can be incorporated into model 2.1 in at least two main ways. Firstly, one could include an error term into the differential equation to give a stochastic differential equation in which the dynamics are partially deterministic but with variability around the deterministic trajectory. Secondly, the dynamics could be simulated as a stochastic Markov Chain. In this

approach, events and the time at which they occur are drawn from a probability distribution (see Section 2.4.1). In this formulation, the system has an ‘absorbing’ state, extinction, from which it cannot recover.

2.2.3 How to model time

The sequencing of events matters for the dynamic processes that underpin population persistence. The decision about how to model time can be thought of as one about whether it should be modelled implicitly as a discrete variable by encoding the sequential nature of events, or explicitly as a continuous variable. Real population processes take place in continuous time in the sense that the interval between events can be arbitrarily small, and in this sense, continuous time models more closely approximate natural processes. This is, for example, the case when there is no *a priori* reason to expect that the process should take place at discrete time intervals, for example, if modelling some cellular processes.

However, there are situations in which modelling in discrete time represents a justifiable simplification. For example, many plants and animals reproduce and die according to seasonal climatic patterns. In the case where the whole population dies at the end of one growing season and is replaced by an entirely new population in the next season - as is the case with annual plants and certain insect species, discrete time modelling is the more appropriate decision. Populations following this pattern are referred to in the theoretical ecological modelling literature as having non-overlapping generations. However, even in the case of species with overlapping generations, a discrete time assumption may be realistic. For example, many species of animals with multiple breeding seasons may still be alive and producing offspring when their offspring begin to reproduce; however, an individual’s offspring does not reproduce during the season in which it was born, and updating the population in discrete time steps is therefore justifiable.

Even in cases where discrete updating is not justified biologically, discrete time may be used to approximate continuous time by selecting a time step that is small in relation to the biological processes under consideration. In this case, a decision needs to be made about the scale on which to aggregate time and thus the events associated with it. Examples of paradigms that model time as discrete are those based on difference equations and cellular automata, whilst approaches based on differential equations and simulation approaches such as the Gillespie algorithm model time as continuous.

2.2.4 How to model the population

A modelling decision that has attracted interest in the literature is that of whether to explicitly model individuals (Durrett and Levin, 1994). Although organisms are discrete, they may be

modelled as continuous, as is the case in Equation 2.1. This assumption is more appropriate for large populations in which individuals have only a small effect on the whole. It may also prove difficult to justify in cases where we are interested in modelling individual-level processes such as extinction, since with continuous population densities, the population can become arbitrarily small without becoming extinct.

A related modelling decision concerns whether to maintain individuality or aggregate the population. In the most disaggregated models, individuals are considered as agents, possibly with their own set of characteristics, and are modelled separately. The paradigmatic examples of this approach are individual-based models and agent-based simulation.

In situations in which individuals fall into relatively clear classes, it might be unnecessary to model each individual separately and we can choose instead to aggregate across individuals. This makes the modelling process more scalable because instead of modelling each individual, we instead keep a count of the number of individuals possessing specific characteristics. For example, social structure models aggregate over individuals, according to age structure or life stage (e.g. by distinguishing between juveniles, and adults capable of reproduction).

Another way in which organisms are often aggregated is in relation to their spatial location. Organisms may be modelled as individuals with a given spatial location, or grouped together in subpopulations at particular spatial locations or on a network. Models that group individuals into subpopulations are known as metapopulation models; the processes that act within and between subpopulations differ. In the simplest case, subpopulations are simply modelled as occupied or unoccupied; in other metapopulation models, individuals are modelled discretely within the subpopulations, but the processes that act within these subpopulations differ from those acting between populations (e.g. because it is less common to come into contact with individuals from different subpopulations than from one's own subpopulation).

Examples of continuous population models are all of those based on differential equations; continuous spatial models often use reaction-diffusion models based on partial differential equations. Discrete population models include lattice, network and metapopulation models.²

Figure 2.1 shows the relationship between modelling paradigms that aggregate differently over individuals and space, organised following our discussion.

²There is lack of consistency in the use of the terms 'patch models' and 'metapopulation models'. We may choose to model organisms at the individual level within patches, or simply model the state of the patches. Throughout this thesis, where the focus is on the latter, we use the term 'metapopulation model' to refer to models in which there is no individuality.

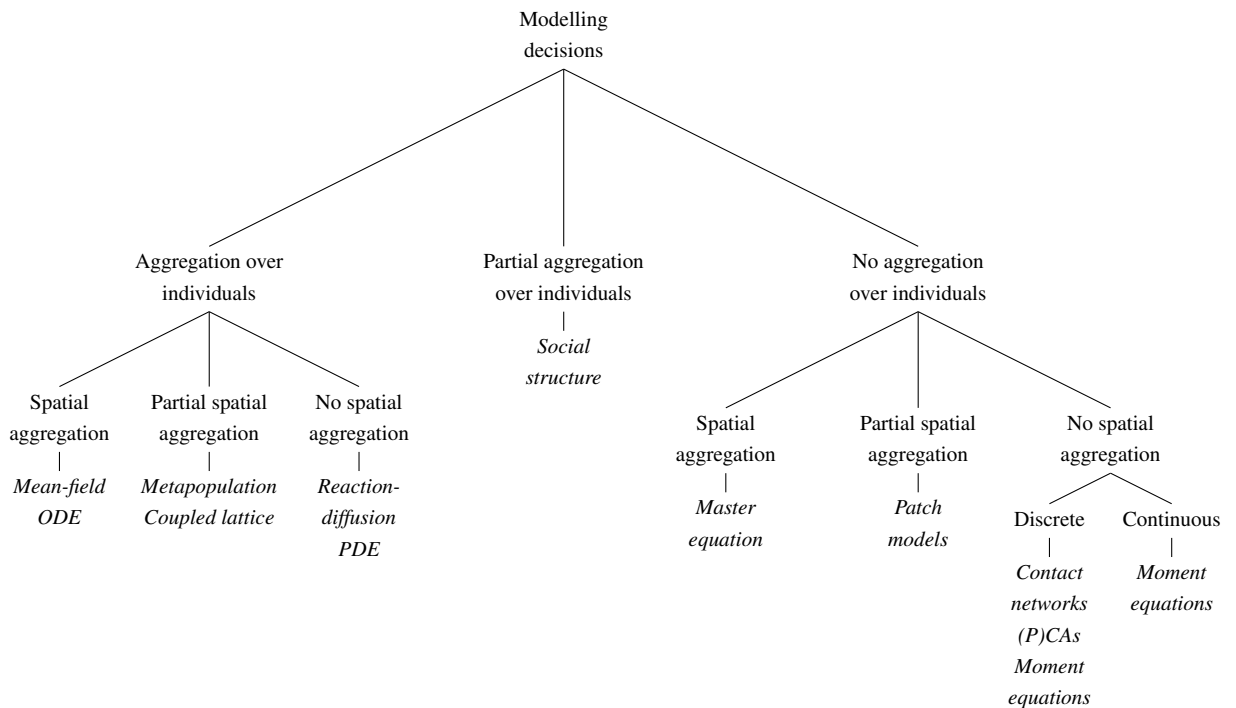


Figure 2.1: Illustration of the relationship between model types resulting from decisions about how to aggregate over individuals and space.

2.2.5 How to model states

An additional decision to be made relates to the states applicable to elements of the system (e.g. organisms or patches). In the site-occupancy interpretation of the logistic model, sites are either occupied or vacant. However, multiple states may be applicable: sites might be occupied by more than one organism in multi-state models, or by organisms in different states such as when they pass through different life stages. Models in epidemiology often demonstrate a larger number of states, and these are explained in more detail in Chapter 6, along with a discussion of possible applications of common epidemiological multi-state models in ecology.

In the example above, individual sites are modelled as being in one of a small number of discrete states. However, in general, individuals themselves may have continuous characteristics. For example, we could model organisms as having a continuous characteristic of age.

2.2.6 Relationships between modelling decisions

Different combinations of modelling decisions lead to different kinds of models. For example, one could choose to implement a metapopulation model with two or three states, in

discrete or continuous time, and do so with or without stochasticity. These modelling decisions are made through consideration of the biological system, the questions one hopes to answer and relevant properties of the models considered, and the intellectual and physical resources (including time, computing power, etc.) available to us to allow us to use the model to understand the biological processes better.

Orzack (2012) argues that focusing on clarifying processes for making modelling decisions in ecology and epidemiology is a useful endeavour. In our framework, we present models according to the modelling decisions regularly taken by those involved in ecological and epidemiological modelling. This approach may be more helpful than simply listing paradigmatic examples because it highlights the similarities and differences between them, while suggesting alternatives. Following the framework above, one would begin by identifying the biological entities that need to be accounted for in a model, and then consider how each of these should be modelled (and indeed whether it should be modelled explicitly at all).

There are several remaining questions that it would be useful to address in relation to the presentation of modelling decisions in the preceding sections, but that are beyond the scope of this thesis. Firstly, it would be useful to explore whether the modelling decisions above can always be made independently of one another, or whether some need to be taken jointly. For example, we might ask how decisions about modelling individuals as continuous or discrete affect decisions about levels of aggregation. Secondly, we provide little information here on how models associated with the different modelling decisions might be implemented, or the kinds of uses (e.g. in the sense of Odenbaugh, 2005) they might support most effectively, and these would also benefit from clarification.

2.2.7 Modelling decisions as a computational biology question

As a relatively young discipline, the question of what constitutes computing science has attracted significant attention and answers have undergone considerable change over the past decades. The debates on this topic typically centre on the distinctions between different research paradigms, often thought of as corresponding to those of natural science, engineering and mathematics, with the most contentious issue that of the extent to which computing science can be considered science (Tedre, 2011). Denning traces these trends (Denning and Freeman, 2009; Denning, 2013), arguing that in the 1950s, a science understanding of the discipline was fairly prevalent, but that an engineering perspective then became more prominent. During the 1990s, a new tendency to reconsider scientific aspects of computing science reemerged as the result of developments in the area of computational science, and the recognition that the notion of computation could also apply to ‘natural’ processes in systems such as DNA (Denning, 2013).

In an attempt to unify the different themes within computing science, Denning (2003) has proposed that the discipline of computing science be characterised by guiding principles. These ‘Great Principles’ consist of principles of design, principles relating to the mechanisms of computing, and practices routinely applied by practitioners. *Design principles* encapsulate conventions followed in order to build good systems, focusing on concepts such as simplicity and performance of computing systems; *mechanics principles* focus on fundamental laws and questions of computation such as computability and system coordination; and *practices* relate to embodied knowledge in the form of habits or routines and include multi-lingual programming skills, engineering practices, modelling and validation and innovation. In an interview with *Ubiquity*, Denning (2007) commented that ‘to be a complete computing professional, you should know the principles and be competent in the four practices.’ Denning claims that it is not through our ability to be explicit about our principles that we should be judged, but by how we conduct our work as professionals. In Appendix A.1, we provide illustrative examples to demonstrate how the principles and practices are applied in the work reported in this thesis, demonstrating how this work meets the criteria for professional work in the discipline³.

A major limitation of the framework is that it does not include any comparison with other disciplines, be they mathematics, engineering or science, although interdisciplinary experience suggests that these practices and principles are indeed more salient (or commitment to them stronger) among those trained in computing science, at least in comparison with those trained in biology and ecology (the contrast with those trained in mathematics appears less stark). The framework also appears to be more appropriate for describing the practices and principles for applications to commercial computing science, as opposed to scientific computing (interpretations of some categories for scientific computing are provided in Appendix A.1 as necessary). Finally, the changes introduced into the framework between 2003 and 2007 appear to add additional complexity, which seems at odds with the aim of providing a coherent structure, and the overall framework appears not to have stabilised.

Moving on from this discussion of the aspects of computing science demonstrated in our work, in the following sections we describe background literature that helps to situate the questions addressed in this research within domain of computing science. We argue that the work described in this thesis contributes to the sub-discipline of computational science⁴, and

³By 2007, the exact content and configuration of the Great Principles had changed, and the later version can be found on Denning’s website at <http://denninginstitute.com/pjd/GP/>. We refer to the 2003 version because of its greater clarity.

⁴Computational science is usually considered to be inherently interdisciplinary; while it can be thought of as the ‘fourth great domain of science’ (Denning and Rosenbloom, 2009) or alternatively the ‘third pillar of science’ (President’s Information Technology Advisory Committee, 2005), discussions along these lines nonetheless consider it as a component of computing science (Denning, 2005), and computational biology (as a discipline of study or research group) is often located within departments of computing science (e.g. at the University of Oxford and Stanford University).

more specifically that of computational biology, using methods from experimental computing science for algorithmic development, either by comparing existing algorithms or developing new algorithms to approximate solutions to intractable problems.

According to the Society for Industrial and Applied Mathematics (SIAM Working Group on CSE Education, 1998), research in computational science and engineering ‘focuses on the development of problem-solving methodologies and robust tools for the solution of scientific and engineering problems’. Another similar definition is provided in the report *Computational Science: Ensuring America’s Competitiveness* (President’s Information Technology Advisory Committee, 2005, p10) which states that ‘at one level, computational science is simply the application of computing capabilities to the solution of problems in the real world’, noting that it uses ‘advanced computing capabilities to understand and solve complex problems’. According to the report, key elements of computational science include algorithms, and modelling and simulation platforms designed to solve scientific problems. For computational science in the context of biology, the National Institute of Health Bioinformatics Definition Committee distinguishes between bioinformatics and computational biology, defining the latter as ‘the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems’ (National Institute of Health - Bioinformatics Definition Committee, 2000).

These definitions appear to focus on research underpinned by scientific questions arising in the natural world (applications of the methods) and engineering questions arising in the development of the methods (methodological developments). Among these, although the engineering questions suggest the development of new methods, work that compares existing computational methods (that are or could be used to answer questions in the natural sciences) also underpins methodological development, and is thus consistent with the general tenor of the definitions. Indeed, as computational methods are applied to new areas, initial testing conducted during the original development phase may prove insufficient and additional testing can be extremely valuable. More explicit acknowledgement of analysis of computational methods in the definitions above would therefore better capture research in the areas of both computational science and computational biology.

The term *experimental computing science* refers to research in computing science (and its sub-disciplines) that employs an experimental methodology based on hypothesis testing (Tedre, 2011), and consists of structured cycles of observation, measurement and analysis (Freeman, 2008). In her article on experimental algorithmics, McGeoch (2007) distinguishes between problems of algorithm analysis and algorithm design: the former is concerned with understanding how different algorithms will perform under given conditions and assumptions; the latter with designing better algorithms, where the term ‘better’ typically implies faster, but may also refer to returning higher quality (more accurate or optimal) solutions

to a given problem. McGeoch argues that experimental approaches to algorithm design and analysis can be highly valuable as a complementary or alternative approach to asymptotic analysis, for example, when we are interested in more than worst case asymptotics in run-time analysis. In the experiments reported in this thesis, we compare candidate alternative algorithms to establish which provides the best ‘explanation’ for (i.e. correspondence with) a baseline algorithm.

Our analysis in this thesis focuses on modelling decisions that are made as a biological model is translated into a form that can be simulated or solved numerically. A useful way to distinguish between different kinds of modelling decisions for scientific computing is via the CoSMoS framework. This framework (Andrews et al., 2011; Stepney and Andrews, 2015) employs key concepts of domain, domain model, platform model, simulation platform and results model. The *domain* refers to the real-world system under consideration; in our context, this is an ecological or epidemiological system. The *domain model* expresses existing understanding (and assumptions) of the biological system at an appropriate level of abstraction for the research questions that will be investigated using the simulation. It is, however, biological: it does not contain any of the details of the implementation of the simulation and is used to separate the science from the implementation details of the simulation platform. The *platform model* is an engineering derivation from the domain model, described in a form that can be easily translated into the technologies of the simulation platform. It should therefore specify engineering design decisions, detailing the structures, behaviours and interactions of the domain model in a form that can be implemented (e.g. for an implementation in Java, in the form of classes, or for an implementation in Matlab in the form of matrix data structures). It also incorporates the necessary parameters to conduct the experimental work, and any implementation constraints and assumptions. The *simulation platform* provides the encoding of the platform model in a way that experiments can be conducted; for example, in Alden et al. (2012), the simulation platform consists of the Java source code of the simulator itself. The platform also includes any auxiliary code required to run the experiments and process the output into a form that can be interpreted. Finally, the *results model* specifies how simulation outputs can be interpreted in relation to the real-world system. Its contents are compared to the domain model to establish the extent to which the simulation output corresponds to that of the real-world domain. The relationship between the domain model and the domain is therefore analogous to that between the simulation platform and the results model.

The work in this thesis focuses on a particular aspect of the CoSMoS framework. Specifically, we take the domain model as given: the lattice logistic model and Spatially Realistic Levins Model are taken directly from the ecological literature. Our interest is in decisions made in order to construct the platform model. However, there are at least two kinds of decisions that can be made at this stage: mathematical assumptions in the form of decisions

that affect the mathematical properties of the platform model and thus potentially the output; and engineering decisions affecting the implementation of the model (in source code) and that influence readability, structure and runtime. Our main interest is in the former kind of design decisions. These may be driven by the constraints of the implementation platform (e.g. hardware constraints such as memory or software constraints such as maximum array sizes) or by less constraining influences such as ease of implementation.

Although we are conscious of the question of engineering decisions in our practice (see Appendix A.1), our primary *research* interest is in the effects of modelling decisions that consist of (or entail) mathematical assumptions that may affect the output of simulations and numerical problem solutions, and thus the scientific conclusions that are drawn from them. Despite the fact that these kinds of decisions can affect scientific conclusions (e.g. Ruxton and Saravia, 1998), they are made in the translation of the domain model to the platform model, and are typically paid little attention in work by practising scientists. As Hogeweg (1988, p88) points out ‘in many modeling efforts little attention is given to choosing a suitable modeling formalism: the choice is most often made on the basis of habit within a certain research area (e.g., in insect population dynamics discrete timestep models are most often used even in the case of overlapping generations, whereas in other population-dynamics areas continuous models abound, even in the case of nonoverlapping generations).’ More recently, Polack (2014), referring to agent-based models for scientific applications, has commented that ‘most reported computer simulations are judged to be valid if they produce something like the expected results’. Indeed, two recent articles, one published in a biological journal (Wilson et al., 2014) and another in a cross-over journal publishing ‘big data’ research in biology (Hastings et al., 2014), concerning best practices for scientific computing, focus almost solely on engineering decisions. The questions of modelling decisions thus appear to be of more interest to applied mathematicians and researchers with strong involvement in computing science (such as Hogeweg and Stepney; see references above). Our specific contributions to the computational biology literature are described in Chapter 7, the main Discussion.

2.3 Persistence

In this section, we discuss notions of persistence that can be found in the literature. We focus primarily on the infectious disease context, while acknowledging a parallel literature on persistence in ecological models. Much of the work on infectious disease epidemiology is centred around understanding epidemics. Work on disease invasion is thus well developed, and theory focuses most notably on the basic reproduction number. The *basic reproduction number* of an infectious disease, denoted R_0 , is defined as the number of secondary infections

arising from a single case, in a fully susceptible population. It is particularly useful for understanding disease invasion: in deterministic models, a disease invades when $R_0 > 1$. In contrast, disease persistence remains relatively under-theorised. A range of definitions of disease persistence and endemicity exist, although the relationship between these is not always clear. In this section, we discuss existing definitions of disease persistence and their connections with related notions in the literature.

In order to structure this discussion, we introduce a compartmental model that describes disease dynamics at the population level, and use this to interpret existing definitions. Although compartmental models are common in the epidemiological literature (as discussed in more detail in Chapter 6), they are typically used to describe the infection state (susceptible, infectious, recovered, etc.) of individuals. In the model proposed here, and illustrated in Figure 2.2, the states describe the status of a disease or pathogen in a given population as absent, epidemic or endemic; the associated processes are those of invasion, establishment and fade-out, with persistence used to describe the process that maintains the pathogen in the endemic state.

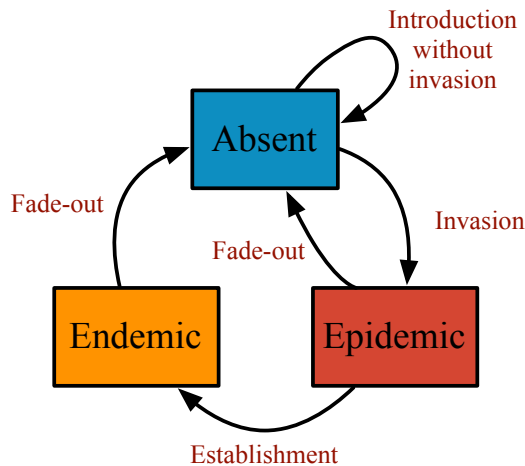


Figure 2.2: Compartmental model of population disease states.

Initially, a pathogen is *absent* from the population. When it is introduced, it either quickly *fades out* without generating an epidemic or *invades* the population (growth in pathogen numbers) and enters the *epidemic* state. If the epidemic does not *fade out* (fall in pathogen numbers to zero) directly from the epidemic state, it *establishes* itself in the population and *persists* in the *endemic* state for some time (pathogen numbers either constant or cyclical). Ultimately, it *fades out* to become once again absent from the population (fall in pathogen numbers to zero).

This discretisation of status into three states is necessarily a simplification; however, it corresponds to the implicit discretisation embedded in much of the language used to describe

population-level disease status (e.g. Nasell, 1995; Castle and Gilligan, 2012). Indeed, although some authors (Gilligan and van den Bosch, 2008) illustrate disease processes in ways that suggest a continuous context (see Figure 2.3 in which, for example, it is not clear at which time point one would consider the disease represented by the yellow line to have entered the endemic state), states are often referred to as discrete.

We now consider definitions of endemicity and persistence. Focusing on verbal descriptions, a range of definitions of endemicity and persistence can be found in the literature. For example, Gilligan and van den Bosch (2008) consider persistence criteria and time to extinction, but appear to use the term *persistence* to refer to both duration of an epidemic and persistence in an endemic state. In another definition, Onstad and Kornkven (1992, p. 561) define endemicity as ‘the persistence or constant presence of a pathogen in an ecologically proper spatial unit over many generations’ while Hagenaars et al. (2004, p. 349) describe persistence as ‘the ability of an infectious disease to avoid extinction’. Both of these definitions focus on the process of persistence once the endemic state is reached, although the timescales over which this might occur are left open. A definition that relates to deterministic modelling is that of Castle and Gilligan (2012), who state that ‘a pathogen is considered to persist if the infection levels reach stable endemic equilibrium values’ (p. 6), a definition that encompasses both establishment and persistence.

Although a number of definitions thus exist, the tendency in most of the literature appears to be to use the language of endemicity and persistence without formally defining it. In addition, verbal definitions are transformed into operational definitions in different ways according to the type of study in which they are employed. The operational definitions can be grouped into three main categories, associated with studies using deterministic models, stochastic models, or a data driven approach.

2.3.1 Operational definitions in deterministic models

Among authors studying deterministic models of disease persistence, the most common approach is to refer to the asymptotic or steady state, or to the endemic equilibrium of a system (Kermack and McKendrick, 1927; Jeger and van den Bosch, 1993; Sun et al., 2012; Li et al., 1999; Post et al., 1983; Roberts, 2007). A similar definition can be found in Cuniffe and Gilligan (2010) who define the ‘endemic level of infection’ as the proportion of infectious individuals as time tends to infinity (i.e. $\lim_{t \rightarrow \infty} I$). The endemic equilibrium is generally found by setting the differential equations governing the disease systems equal to zero, and then solving either analytically or numerically and focusing on non-zero stationary solutions; where this is impossible, authors may simply consider whether a non-zero equilibrium exists. In other words, an endemic situation is operationalised as one in which the model has a non-trivial equilibrium solution.

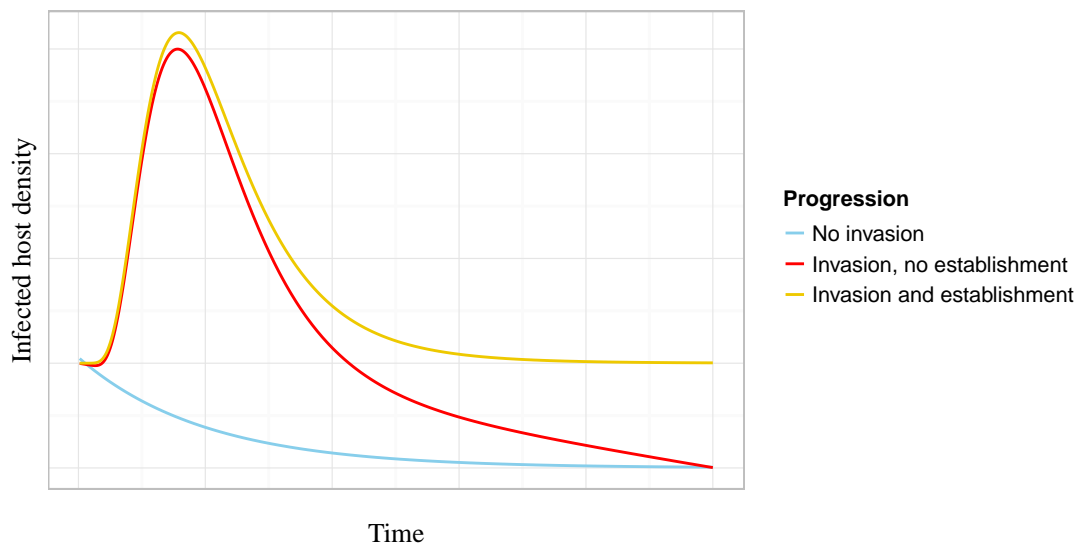


Figure 2.3: Conceptual illustration of three scenarios following introduction of an infectious agent. Redrawn from Gubbins et al. (2000); Gilligan and van den Bosch (2008).

2.3.2 Operational definitions in stochastic models

As explained in Nasell (1995), no steady state exists in (stochastic) Markov models as extinction is certain in the long run. Research using stochastic models therefore tends to operationalise endemicity in relation to persistence beyond predefined time points. For example, Onstad and Kornkven (1992) consider that a pathogen persists if it is still present after 1000 time units. In more applied work, time windows may be chosen according to known characteristics of target populations. For example, Keeling and Gilligan (2000) also use a fixed time window to differentiate between endemic persistence and epidemics in a stochastic model of bubonic plague in rodent populations, with endemic persistence defined as post-introduction persistence beyond 10 years. Thus, although they differentiate verbally between invasion and persistence, this information is not included in the operational definition. Hagenaars et al. (2004) allow external re-infection, and take a slightly different approach, operationalising persistence in the form of the *fade-out fraction*, defined as the ‘expected proportion of time that the disease is extinct in a system’ (Hagenaars et al., 2004, p351). This definition allows more than a binary distinction between realisations that went extinct and those that did not; however, it relies on an external source of infection with known frequency of re-infections.

More generally, setting a specific time window only allows us to distinguish between realisations in which extinction occurred and those in which it did not (although over multiple realisations, proportions can be calculated). In fact, even Hagenaars et al.’s (2004) definition relies on a capture window. The time window approaches also fail to capture the distinction between instantiations that were near extinction at the time when persistence is evaluated; if

a falling population is considered to be a sign of having left the endemic state (in line with deterministic definitions), such approaches overestimate persistence in the endemic state.

Other measures of persistence have been considered in the literature. For example, in their simulation study, Jesse and Heesterbeek (2011, p14) considered ‘the infectious agent to be persistent in a simulation if it is still present in the population after twice the expected life span of the host’ but conceded that ‘the choice of twice the expected life span is arbitrary’. They justify their choice by noting that ‘within this number of time steps the infectious agent has survived two generations of hosts, and has spread between the patches’.

A more mathematically rigorous alternative would be to consider a disease to be endemic in the case where the quasi-stationary distribution (QSD) is reached. The quasi-stationary distribution is the distribution of systems states, conditioned on non-extinction. It is defined formally in Section 2.4.2 and explored in more depth in later sections, especially Chapter 5. Nasell (2005, p. 204) argues that it is the ‘counterpart to the endemic infection level in the deterministic model’. It therefore provides a corresponding definition in the stochastic context to that of Castle and Gilligan’s (2012) definition for deterministic systems.

Using the notion of the QSD is helpful for a number of reasons. Firstly, once a system has entered the QSD, extinction times are known to be exponentially distributed (Artalejo, 2012). This means that the full distribution of extinction times is captured by a single parameter, the mean time until extinction, usually denoted T_m . The characteristics of this distribution mean that $1/T_m$ gives the rate at which populations go extinct, a quantity that may be easier to measure in real populations since it does not rely on the extinction of all comparable populations.

2.3.3 Operational definitions in data-driven work

Examples of operationalisations of endemicity in empirically-driven work tend to focus on levels of infection. In addition to definitions that allow us to identify whether or not a system is in an endemic state, another question of interest is that of the proportion of the host population that is infected (either symptomatically or asymptotically) in the endemic state. For example, Hay et al. (2009) follow earlier work in considering three classes of endemicity and three levels of malaria endemicity within the stable risk class. Levels of endemicity within stable risk areas were categorised as low (less than 5% of the population with blood-detectable parasites), intermediate (between 5% and 40%) and high (over 40%). However, the split between classes appears rather arbitrary.

2.3.4 Other related definitions

Similar definitions have been discussed in the literature on critical community size and disease reservoirs. Bartlett (1957a) introduced the notion of critical community size to capture the idea that larger host populations are more able to support the persistence of pathogens than smaller populations. *Critical community size* was defined as the population size ‘for which the chance of fade-out after a major epidemic is 50 per cent’ (Bartlett, 1957a, p. 56), thus referring to the host population threshold above which the chance of fade-out was below 50%. This definition appears to encompass the processes of both establishment and persistence, while setting a threshold for persistence. However, the notion of critical community size is problematic both conceptually and because it is difficult to operationalise (Viana et al., 2014). From a conceptual point of view, the notion of population size fails to capture other epidemiological characteristics of relevance such as connectedness of subpopulations; furthermore, the terms ‘fade-out’ and ‘major’ are vague; finally, while the definition employs the point of 50% probability of persistence, the distribution around persistence probabilities remains unspecified. It is not clear how to link the notion of a critical community to disease outcomes of interest: for example, it is unclear why we should be more interested in a 50% chance of ‘fade-out’ than, say, persistence until such time as there is an X% risk of transmission to the target population. Further, measuring critical community size is, by definition, problematic since it requires several instances of fade-out in communities of different sizes. For many diseases, conducting such measurements is unrealistic.

In more recent work on critical community size, Nasell (2005, p. 210) formalised Bartlett’s definition such that the critical size is defined as ‘that value of N for which the probability of extinction after waiting for one quasi-period T_0 equals 0.5’. In the case of Bartlett’s (1957a) definition, the starting point for the waiting times is after a ‘major epidemic’ whereas Nasell (2005) defines the waiting times as starting from an initial distribution matching that of the quasi-stationary distribution. In the context of this discussion, Nasell’s (2005) definition therefore focuses on persistence once in the endemic state.

Other related definitions can be found in the literature on disease reservoirs. For example, Ashford (1997, 2003, p1495) proposed to define a reservoir as an ‘ecological system in which the infectious agent survives indefinitely’, whereas Haydon et al. (2002, p1469) defined a reservoir as ‘one or more epidemiologically connected populations or environments in which the pathogen can be permanently maintained and from which infection is transmitted to the defined target population’. According to the first of these definitions, a reservoir could therefore be defined as an ecological system in which a pathogen is endemic; in the second, the endemic status of the reservoir is the same, but the reservoir is defined in relation to a specified target population of interest (see also Viana et al., 2014).

2.3.5 Persistence and model types

The discussion in the preceding section demonstrates that although definitions of persistence and endemicity exist, they are made operational in a range of ways. Some of these relate to whether processes are considered to act deterministically or stochastically. In the deterministic case, persistence is a relatively simple concept implied by the existence of a stable non-trivial equilibrium. However, if we believe that real-world systems are stochastic, this means that this definition cannot be instantiated in the real world. In the stochastic case, different ways of operationalising persistence have emerged, focusing on the existence of a population until a particular time horizon, the time to extinction of a population, the critical size of a community required to support a population that persists with a given probability (usually 50%). These different operationalisations make comparisons between different studies difficult. For example, if a particular paper reports 50% persistence until a time horizon of 100 years, and another paper persistence until 25 years, there is no way to compare these findings.

In addition to operational differences, there is little consensus on use of terminology. Some authors use the term *persistence* to refer to the process from the point of initial introduction to extinction, whereas others use it to refer to the process only once the endemic phase is reached. This makes it particularly important to interrogate the operational definitions used when comparing the findings from different studies.

It seems likely that work on persistence would benefit from a more unified approach in which terminology is used more consistently, and operationalisations of persistence are more clearly delineated. A better mapping between definitions of persistence and meaningful disease outcomes of relevance to practitioners and those working to control disease would clearly be beneficial. The framework proposed in Figure 2.2 provides a partial resolution to these issues through clarifying the terminology. According to this framework, the term *persistence* should be reserved for describing the situation of having reached a stable state in a deterministic model, or the process of continued existence of a population once the QSD has been reached. Similarly, referring to an organism as being *endemic* therefore corresponds to it having reached a stable state in deterministic models or in the stochastic case, of being in the quasi-stationary phase.

The framework formalises the states and transitions discussed in the literature and clarifies the distinctions between them, while providing a standardised language for discussing these ideas. This should be beneficial for the comparison of studies of persistence, as well as in making modelling decisions about how to choose an appropriate definition. Additional questions relate to whether there are situations in which different definitions of persistence can be considered comparable, and how to re-interpret data from existing studies in such a way that they can be compared. It would also be valuable to consider more explicitly the

appropriateness of specific definitions for particular aims.

2.4 Probability theory and simulation concepts

Two key probability theory ideas are applied several times in the thesis, and these are now introduced. The first of these is the Gillespie algorithm; the second is the quasi-stationary distribution.

2.4.1 Gillespie algorithm

Much of the work in this thesis is concerned with stochastic simulation in continuous time. These simulations rely on the Gillespie algorithm and variants of it, as applied to spatially explicit systems. More specifically, they all use variants of the Direct Method described in Gillespie (1977), used to generate a statistically correct trajectory of a stochastic equation. Based in probability theory, this algorithm was originally made popular in its use for the simulation of chemical and biochemical reactions, and there remains a strong interest in its use and improvements in the biochemical and chemical literature (Sanassy et al., 2014); its use is also common in the ecological literature (Black and McKane, 2012).

The simulation algorithm employs the average rates of reactions or events that can take place in the system. In the context of an ecological system, these might be birth or death events; in the context of metapopulations, they are colonisation and extinction events. The algorithm uses properties of the exponential distribution which allow us to sample separately the time until the next event and the event that occurs.

We denote the events that could occur (e.g. births and deaths) by E_m , up to a total of E_M events. If organisms have the same event rates, the number of organisms that could undergo each event is denoted h_m , while the rate at which each organism undergoes this event is denoted c_m . Note that if organisms have different event rates - either because they have different inherent rates or because of ecological effects such as overcrowding - then we typically allocate a separate birth and death rate to each organism and $h_m = 1$ for all events. We sum over each of the M events to calculate the total event rate rate $E_0 = \sum_{m=1}^M c_m h_m$.

The algorithm is described procedurally in Algorithm 2.1. The first step is to sample the time until the next event τ . This follows an exponential distribution with rate parameter E_0 , and can be sampled from this distribution⁵. The next step is to decide which event should occur. For this, we need to work out the probabilities of the different events, which we do by dividing by the total rate. We then select an event probabilistically, in proportion to

⁵This can be achieved by drawing a number r_1 from the uniform distribution on the interval $(0, 1)$ and calculating $\tau = (1/E_0)\ln(1/r_1)$.

its probability. We can think of the event probabilities as covering the unit interval, and generate a random number r_2 on the interval $[0, 1)$ and use it to choose the event ν that spans the section of the interval containing r_2 .

Algorithm 2.1 Gillespie algorithm, direct method

- 1: **while** $t < t_{max}$ **do**
 - 2: Compute event rates E_m and total event rate E_0
 - 3: Generate $r_1, r_2 \sim U(0, 1)$
 - 4: Compute time to next event $\tau = (1/E_0)\ln(1/r_1)$
 - 5: Update time: $t \leftarrow t + \tau$
 - 6: Choose event E_ν for which $\sum_{m=1}^{M-1} E_m/E_0 < r_2 \leq \sum_{m=1}^M E_m/E_0$
 - 7: Update system state: perform event E_ν
 - 8: **end while**
-

More computationally efficient algorithms have been developed; these are reviewed in Sanassy et al. (2014). For example, although the original algorithm used a linear search to select the next event (line 6), improvements use more efficient search algorithms with better scaling properties. In addition to these improvements to the exact algorithm, approximate methods have been developed. The most commonly used approximation is that of tau-leaping (or τ -leaping; Gillespie, 2001). In this approximation, instead of simulating each event individually and updating rates after each event, we simulate together all of the events expected to take place within the time interval $[t, t + \tau)$. The process is shown in Algorithm 2.2. Because the number K_m is unbounded, it is necessary to check that unrealistic (or negative) values are not reached before conducting the updating step. The approximation is appropriate when the state of the system does not change too much during the time interval τ . A method for efficient step size selection is explained in Cao et al. (2006).

Algorithm 2.2 Gillespie algorithm, tau-leaping approximation

- 1: **while** $t < t_{max}$ **do**
 - 2: Compute event rates E_m
 - 3: Choose a time step τ
 - 4: For each event E_m , generate an event count $K_m \sim \text{Poisson}(E_m\tau)$
 - 5: Update system state: perform each event E_m a number K_m times
 - 6: Update time: $t \leftarrow t + \tau$
 - 7: **end while**
-

2.4.2 The quasi-stationary distribution

The concept of the quasi-stationary distribution is usually invoked in systems modelled as Markov chains with at least one absorbing state. In the context of population dynamics, a Markov chain might be used to model the number of organisms in a system, and the absorbing state would correspond to the extinction state of the species. Intuitively, the quasi-stationary distribution describes the long-run distribution of system states (e.g. patch occupancy states or number of individuals present), prior to extinction.

Pollett (2012) explains that the idea of the quasi-stationary distribution (or conditional limiting distribution) can be traced back to Yaglom (1947). However, he notes that the idea actually had deeper roots in ecology and evolution, noting that Wright (1931) had already referred to a limiting conditional distribution of gene frequencies. A very early use of the term *quasi-stationarity* can also be found in the epidemiological literature in work by Bartlett (1956, 1957b), who referred to an ‘an effective or quasi-stationarity’ (Bartlett, 1957b, p.38), and according to Pollett (2012), coined the term *quasi-stationary distribution* in his 1960 work (Bartlett, 1960). A formal definition and general theory became available in the early 1960s (Pollett, 2012), with a continuous-time version of the theory developed in Darroch and Seneta (1967).

We now define the QSD more formally, employing a combination of the notation used in Artalejo (2012) and de Oliveira and Dickman (2005). Specifically, we define a regular time-homogeneous Markov process $X = \{X(t); t \geq 0\}$ on a countable state space S (in the finite case, this is of dimension m). The state space S consists of a set S_T of transient states among which the process evolves until it hits a set of one or more absorbing states S_A ; thus $S = S_T \cup S_A$. Without loss of generality, the states can be labelled such that $S_A = \{0\}$, and in the finite state case, the transitory states are labelled $S_T = \{1, 2, \dots, m\}$. In other words, the Markov process takes the values $\sigma = 0, 1, 2, \dots, m$, with the state $\sigma = 0$ absorbing.

In a first definition, we allow time to tend to infinity and consider the probability of the chain being in a particular state, conditioned on non-extinction. Note that the distribution defined in this way is also referred to as the quasi-limiting distribution (QLD) (see e.g. Méléard and Villemonais, 2012). We use $p_\sigma(t)$ to denote the probability that $X(t) = \sigma$ for some σ in the transient set $\sigma \in S_T$, given some particular initial state $X(0)$ (i.e. $p_\sigma(t) = P\{X(t) = \sigma \in S_T \mid X(0) = j \in S_T\}$). The survival probability $P_T(t) = \sum_{\sigma \geq 1} p_\sigma(t) = 1 - p_0(t)$ is the probability that the process has not become trapped in the absorbing state by time t (i.e. is still in the transient set). Allowing t to tend to infinity, if it exists, the probability density distribution given by the vector π in which $\pi_\sigma = \lim_{t \rightarrow \infty} \frac{p_\sigma(t)}{P_T(t)}$ is a QSD.

In a second definition, we start with a probability distribution for the states of the Markov chain and run the chain forwards. The initial probability distribution is a quasi-stationary distribution if for all future time until extinction, the probability distribution of states in

which the chain finds itself is unchanged (i.e. is time independent). Formally, we denote the time to extinction by $T = \inf\{t > 0 \mid X(t) \in S_A, X(0) \in S_T\}$, corresponding to the lower bound on the time that the process enters the absorbing state. Suppose that $\mathbf{a} = \{a_\sigma; \sigma \in S_T\}$ denotes a probability distribution over the transient states defined by $a_\sigma = P\{X(t) = \sigma\}$. Now, if there exists an initial distribution $\boldsymbol{\pi}$ defined as $\pi_\sigma = P\{X(0) = \sigma\}$, such that for all transient states $\sigma \in S_T$ and for all future times $t \geq 0$ until extinction at time T , $P\{X(t) = \sigma \mid T > t\} = \pi_\sigma$, then $\boldsymbol{\pi}$ is called a quasi-stationary distribution.

In the case where the state space is finite, the two definitions are equivalent (Vere-Jones, 1969; Méléard and Villemonais, 2012). Furthermore, when the chain is finite and irreducible (i.e. every transient state can be reached from every other state), the existence of a unique QSD is guaranteed (Darroch and Seneta, 1967). These conditions are fulfilled by the models considered in this thesis.

Computing the QSD

The standard approach for obtaining the QSD is based on eigenvector analysis. This approach is provably exact up to the limits of numerical approximation in obtaining eigenvalues. The QSD is found numerically from the appropriate generator matrix (see Nasell, 2001b, for a clear explanation of why this is the case). For the full chain (including extinction), the generator matrix \mathbf{Q} in the discrete time case consists of the state transition probabilities matrix in which q_{rs} is the transition probability from state r to state s . In the continuous time case, the transition rate matrix or (infinitesimal) generator matrix \mathbf{Q} is based upon the transition rates: off-diagonal elements q_{rs} are given by the transition rates from r to s while diagonal elements are defined such that row sums are zero, i.e. $q_{rr} = -\sum_{s \neq r} q_{rs}$. For the SRLM, defining $\nu = 2^n$ for notational simplicity, \mathbf{Q} has dimension $\nu \times \nu$. The generator matrix for the continuous case where the zero state is absorbing (rates out of this state are zero) is

$$\mathbf{Q} = \begin{bmatrix} -\sum_{s \neq 1} q_{0s} & 0 & 0 & \cdots & 0 \\ q_{10} & -\sum_{s \neq 1} q_{1s} & q_{12} & \cdots & q_{1\nu} \\ q_{20} & q_{21} & -\sum_{s \neq 2} q_{2s} & \cdots & q_{2\nu} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{\nu 0} & q_{\nu 1} & q_{\nu 2} & \cdots & -\sum_{s \neq \nu} q_{\nu s} \end{bmatrix}$$

In order to obtain the quasi-stationary distribution, we employ the sub-matrix of the generator matrix corresponding to the transient states S_T . The sub-generator matrix \mathbf{Q}_{S_T} is obtained from the full generator matrix by removing the row and column corresponding to transitions

in and out of the absorbing state and has dimension $(\nu - 1) \times (\nu - 1)$:

$$\mathbf{Q}_{S_T} = \begin{bmatrix} -\sum_{s \neq 1} q_{1s} & q_{12} & \cdots & q_{1\nu} \\ q_{21} & -\sum_{s \neq 2} q_{2s} & \cdots & q_{2\nu} \\ \vdots & \vdots & \ddots & \vdots \\ q_{\nu 1} & q_{\nu 2} & \cdots & -\sum_{s \neq \nu} q_{\nu s} \end{bmatrix}$$

The QSD is then obtained by calculating the left eigenvector of \mathbf{Q}_{S_T} corresponding to the most non-negative real eigenvalue and by dividing each element by the sum of the elements so that the entries sum to one⁶.

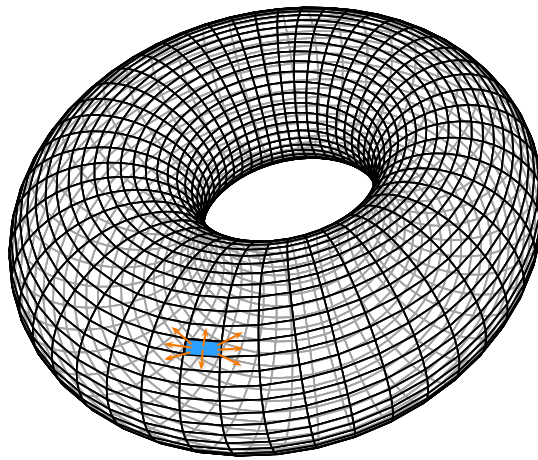
2.5 Models and persistence definitions in this thesis

Two main theoretical domain models, drawn from the ecological literature, are employed in this thesis. These are illustrated in Figure 2.4. In Chapter 3, we use a cellular automaton model on a torus, in which organisms live on a regular grid of sites and can die, or give birth into any one of the eight neighbouring sites (as indicated by the ‘live’ blue cell and arrows). In Chapters 4, 5 and 6, we use that Spatially Realistic Levins Model, a sample landscape for which is illustrated in Figure 2.4 (b). In this model, a landscape is made up of patches (representing areas of habitat, or populations of hosts). Each patch can become colonised by organisms from other patches, or go extinct.

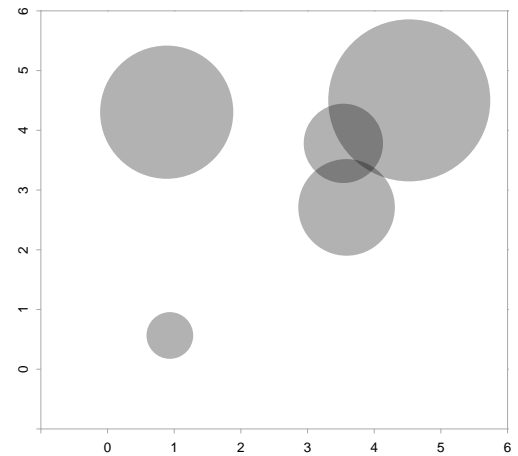
We now provide an overview of the models and persistence definitions used in the studies included in this thesis, summarised in Table 2.1. Firstly, we note that throughout this thesis, we assume that the real world is stochastic, that biological processes occur in continuous time, that individual organisms are discrete and diverse, and that their interactions are spatially heterogeneous. We consider models that embody these assumptions as providing the best representation of real world systems; we then test the effects of changing these and other modelling decisions on the variables of interest.

Three main theoretical models are used in this thesis: a lattice logistic model in Chapter 3; the spatially-realised Levins model in Chapters 4, 5 and 6; and the original Levins model in 6. In relation to the modelling decisions made, we use stochastic implementations of these models in Chapters 3, 4 and 5, where they are implemented as a Markov chain, and deterministic models in Chapter 6. Time is modelled as continuous in all chapters, and we also compare with discrete-time models in Chapter 3. In all chapters, the population status is modelled in relation to pre-specified sites or patches. In Chapter 3 we think of each patch

⁶The ordering is often ambiguous in the literature: the generator matrix must first be constructed and the sub-generator obtained from it; this guarantees the correct values on the diagonal.



(a) Cellular automaton on a torus.



(b) Sample landscape for Spatially Realistic Levins Model (SRLM).

Figure 2.4: Visual representation of models used in the thesis

Table 2.1: Overview of questions and definitions of persistence used in studies described in the thesis.

Chapter	Persistence questions or problems addressed	Model	Definition of persistence
Chapter 3	How are persistence and co-existence affected by decisions to model under continuous versus discrete time?	Lattice logistic model, stochastic, simulations using cellular automaton on torus	Persistence until pre-specified time (1000 time units)
Chapter 4	How to quantify the importance of a patch to persistence in a heterogeneous metapopulation?	Spatially Realistic Levins Model (basic and general version), stochastic, numerical	Mean time to extinction from quasi-stationarity, metapopulation size
Chapter 5	How does compressing the QSD for large, heterogeneous systems affect estimates of persistence?	SRLM, stochastic, simulation	Mean time to extinction from quasi-stationarity
Chapter 6	How does the introduction of additional patch states influence persistence thresholds in metapopulation models?	Levins model (with and without Allee effects), SRLM, deterministic, analytic	Existence of a stable, nontrivial solution

as hosting a single individual; in Chapters 4, 5 and 6 patches host whole populations, but the internal dynamics of the population are not modelled (i.e. we aggregate over individuals and partially aggregate over space). In Chapters 4 and 5 we consider only two states per patch; in Chapter 3, patches can be inhabited by either a single species, or one of two species; in Chapter 6, a larger number of patch states is considered.

In relation to the definitions of persistence, in Chapter 3, we use a simple time horizon approach. In Chapters 4 and 5 we are concerned with the QSD and measure mean time to extinction from quasi-stationarity and in Chapter 4, also consider metapopulation size. In Chapter 6, we work with deterministic models and consider persistence in relation to the existence of a stable non-trivial equilibrium.

2.6 Conclusion

In this chapter, we have presented two frameworks for describing modelling decisions and definitions of persistence. The first constitutes an attempt to classify ecological and epidemiological models in a way that moves beyond the usual approach of presenting a range of model types by describing paradigmatic examples. Instead, we present modelling approaches according to the decisions made about how to represent relevant aspects of the world by those involved in ecological and epidemiological modelling. It might, therefore, be helpful in guiding modelling decisions and their reporting. We also consider how understanding modelling decisions fits within the computational biology literature, and explain that attempts to understand the effects of modelling decisions made during the translation from the (biological) *domain model* to the *platform model* form the basis for legitimate computing science research questions. The second framework, shown visually in Figure 2.2, helps to clarify the language used in work on population persistence, and might be used to provide a standardised terminology for relating persistence concepts found in the literature. In the final sections of this chapter, we introduced two probability theory concepts that are used repeatedly in the thesis - the Gillespie algorithm and the quasi-stationary distribution - before providing an overview of the ways in which all of the concepts discussed in the chapter relate to the studies in the following sections.

Chapter 3

Discrete and continuous time simulations of spatial ecological processes predict different final population sizes and interspecific competition outcomes

Cellular automata (CAs) are commonly used to simulate spatial processes in ecology. Although appropriate for modelling events that occur at discrete time points, they are also routinely used to model biological processes that take place continuously. We report on a study comparing predictions of discrete time CA models to those of their continuous time counterpart. Specifically, we investigate how the decision to model time discretely or continuously affects predictions regarding long-run population sizes, the probability of extinction and interspecific competition. We show effects on predicted ecological outcomes, finding quantitative differences in all cases and in the case of interspecific competition, additional qualitative differences in predictions regarding species dominance. Our findings demonstrate that qualitative conclusions drawn from spatial simulations can be critically dependent on the decision to model time discretely or continuously. Contrary to our expectations, simulating in continuous time did not incur a heavy computational penalty. We also raise ecological questions on the relative benefits of reproductive strategies that take place in discrete and continuous time.¹

¹The original publication, Mancy et al. (2013), can be obtained at <http://dx.doi.org/10.1016/j.ecolmodel.2013.03.013>

3.1 Introduction

Cellular automata (CAs) are commonly used to simulate dynamic spatial processes in ecology, contributing to developments in both applied and theoretical research. In a simple CA model of birth-death processes, individuals inhabit discrete sites, usually organised in a grid formation. Time progresses in discrete steps and an *update scheme* specifies how individuals die or give birth into neighbouring sites at each step. For example, in the applied literature CAs have been used to simulate the spatial distribution of insect colonies (Perfecto and Vandermeer, 2008; Vandermeer et al., 2008) and the effect of plant-soil feedbacks on relative tree abundance (Mangan et al., 2010), while in microbial ecology, Fox et al. (2008) used a CA to investigate the way in which plasmids invade bacterial populations. In contrast, Laird and Schamp (2008) used a CA to explore theoretical questions relating to differences between interspecific competition in spatial and non-spatial (homogeneous mixing) contexts while Roxburgh et al. (2004) investigated mechanisms leading to long-term species coexistence in the context of the intermediate disturbance hypothesis. In addition, CAs have been used to validate simplifications required to solve models analytically, such as in the theoretical work on population self-structuring reviewed in Lion and van Baalen (2008), as well as in a range of publications on the evolution host-parasite interactions (see e.g. Kamo and Boots, 2004; Best et al., 2011, for parasite virulence and host resistance respectively), the evolution of altruism (e.g. Lion and van Baalen, 2007) and the evolution of reproductive effort (Lion, 2010).

Among the assumptions embodied in CA models is that of discrete time. This in turn introduces the need to make additional assumptions in the form of modelling decisions regarding the *update scheme* used to govern the order in which sites are considered and events take place. When these modelling decisions are made carefully, CAs can form appropriate models for discrete time spatial processes. However, they are often employed to simulate continuous time ecological processes or models, frequently without acknowledgement that this introduces an additional layer of approximation. Fortunately, these continuous processes can be simulated directly using a discrete space version of the Gillespie algorithm (Gillespie, 1977)². Following this algorithm, time is continuous in the sense that it progresses in arbitrarily small steps, the length of which varies according to event rates, and these are limited only by the precision of the computer on which it is implemented. Although a little more mathematically involved than discrete time approaches, the implementation of this algorithm reduces the number of modelling decisions and thus allows a stronger focus on the biology while enhancing comparability between studies. Once understood, the approach can also be applied to non-spatial and continuous space problems, as well as to evolutionary problems

²The Gillespie algorithm and the τ -leaping approximation referred to in the following paragraphs are described in Section 2.4.1.

(see e.g. Meier et al., 2011).

Decisions about whether to simulate in continuous or discrete time, and in the latter case which update scheme to use, are not simply technical but should be made in direct relation to the dynamics of the biological system under study. In order to compare studies and make informed decisions about which approach to use, it is important to understand any disparities in predictions between continuous and discrete time simulations, especially in the case where CAs are used to model continuous time processes. Although it is known that CA update schemes (the order in which events are considered) can affect ecological dynamics (Ruxton and Saravia, 1998), differences in ecologically meaningful predictions between CAs and corresponding continuous time simulation approaches have never been tested.

In this chapter, we assume stochastic real world processes that occur in continuous time with exponentially distributed waiting times between events, taking as a case study the asymmetric logistic model of population growth on a lattice (Matsuda et al., 1992). We regard this model as our benchmark and consider discrete time CA simulations as approximations to this model. Specifically, we simulate this model stochastically in continuous time and compare outcomes with those of simulations conducted using two probabilistic CA update schemes, a range of time step sizes and two methods for converting between the rates used in continuous time models and probabilities required for discrete time simulation. We conduct two experiments, focusing in the first on a single species and in the second on competition between two species. In both experiments, we report long-run population sizes, and in the context of interspecific competition, also predictions regarding coexistence and competitive exclusion, using these outcomes to highlight disparities between discrete and continuous time.

In the following sections we consider some of the modelling decisions that need to be made when using CAs, emphasising the conversion from rates to probabilities required to approximate continuous processes. We describe our experimental protocol, provide findings from our two experiments and conclude with modelling recommendations.

3.2 Modelling decisions of cellular automata

Provided that time steps are chosen carefully so that they match the periodicity of real ecological events, discrete time simulations can be appropriate when simulating ecological processes that occur synchronously (e.g. reproductive cycles in cicadas) or where there are strong cyclical patterns (e.g. due to seasonality). Their use becomes more difficult to justify when modelling continuous processes (e.g. disease transmission) or to validate analytic simplifications in continuous time models. Nonetheless, justifications for employing discrete time simulations, decisions regarding particular choices of update scheme and method of converting rates to probabilities are rarely reported (although see Best et al., 2011;

Ovaskainen and Hanski, 2003a, for articles including this information). This makes replication almost impossible, as well as hindering comparisons between studies and the interpretation of any conflicting findings. In this study, we investigate the extent of these problems by simulating the same system in continuous and discrete time, taking a continuous time model as our benchmark. We limit our discussion to one or more species living on a finite grid, and assume ecological processes that take place continuously according to a well-understood model.

Simple CA models of ecological processes are usually constructed in the following way: organisms live on a grid of sites; time progresses in discrete time steps and at each iteration individuals persist, die, or give birth into *neighbouring* sites according to a set of *local transition rules*. In probabilistic models, event probabilities are often dependent on the configuration of occupied and empty sites in the neighbourhood. CAs are relatively straightforward to implement, requiring limited mathematical or modelling knowledge (Berec, 2002; Breckling et al., 2011) but their very ease of implementation belies a range of complexities. Specifically, important modelling decisions arise as a result of the discrete nature of time: these concern the order in which events are executed and the way in which event rates are converted to probabilities.

The issue of event ordering arises because in discrete time, events may occur simultaneously at the same site (e.g. two births into the same site) and decisions thus need to be made about the order in which events should take place and how to resolve competition. An *update scheme* is therefore used to determine event ordering. A large number of schemes have been proposed and comparisons between these in the computing science, theoretical physics and ecological literature demonstrate important differences in dynamics and steady state outcomes (e.g. Manzoni, 2012; Ingerson and Buvel, 1984; Lumer and Nicolis, 1994; Schönfisch and de Roos, 1999; Cornforth et al., 2002; Ruxton and Saravia, 1998).

Event frequency in continuous time is typically characterised by event rates and the assumption that waiting times between events are exponentially distributed. For use in CA models, these rates must be converted into event probabilities. We use two different approaches in our study, one that allows for multiple events and one that allows only a single event per time step Δ_t . In the first, we make use of the fact that for a process with exponentially distributed waiting times, the number of events within a specified time window follows a Poisson distribution. Thus, we sample the number of events from a Poisson distribution with parameter $r\Delta_t$ where r is the instantaneous rate (note that this only makes sense for births). When discrete time is viewed as an approximation to a continuous process, this is similar to the τ -leaping idea proposed by Gillespie (2001). The second conversion is a cruder approximation that allows a maximum of one event per time step, bringing the simulation into line with most common CA approaches (see e.g. Best et al., 2011, for a study where this conversion is described explicitly). Probabilities in this approach are computed from rates as described in

Section 3.3.3. System dynamics are expected to differ between conversion approaches and although it is known that reducing the time step should limit this effect (see e.g. Schönfisch and de Roos, 1999), it is unclear how small Δ_t needs to be before particular (qualitative and quantitative) properties of ecological models are indistinguishable.

3.3 Experimental protocol

Following Ruxton and Saravia's (1998) comparison of CA update schemes, we take as a case study one of the simplest spatial models, the asymmetric logistic model of population growth on a lattice. We simulate a stochastic version of this model for different birth and death rates using a model with exponentially distributed waiting times between events. This is compared to simulations using two CA update schemes, a range of time steps and two methods used to convert from rates to probabilities. In our analysis, we consider the continuous time simulation as our benchmark and the discrete time simulations as approximations to this. In Experiment 1, we simulate a single species and consider differences in long-run population sizes and probability of extinction. In Experiment 2, we consider two species and compare interspecific competition outcomes. Simulations are conducted for 1000 time units and 100 repetitions unless otherwise stated.

3.3.1 Continuous time model

The lattice logistic model describes population growth that is regulated by the local availability of empty sites. In the standard version of the model, organisms live on an infinite network of sites, each of which is connected to n randomly-selected neighbours. Organisms have two fundamental behaviours - birth and death - governed by rates, and can only give birth if there is an empty site in their neighbourhood (Matsuda et al., 1992)³. We simulate the stochastic version of this continuous time model as a Poisson process. This is implemented in the form of a spatial version of the Gillespie algorithm (Gillespie, 1977) that resembles the algorithm proposed by Stundzia and Lumsden (1996) and that we refer to as *Gill*. The algorithms used are described formally in Appendix B.1 and the code is available from (<http://rebeccamancy.github.io/gillespie-cellular-automaton/>).

³In some versions of the lattice logistic model, death rate is also related to the density of neighbours. In the version we implement, death occurs at a constant rate independent of overcrowding while reproduction is limited by resource constraints.

3.3.2 Cellular automaton update schemes

We compare the outcomes of our continuous time simulations to those of a range of CAs. We select two update schemes among those considered by Ruxton and Saravia (1998), deliberately choosing schemes in which site order is random and that differ in the level of structure introduced (Schönfisch and de Roos, 1999) to investigate the effects of update timing and event orderings.

The first scheme introduces structure by fixing the order of birth and death events and through delayed updating. At each generation, a new random site order is generated which is then used to execute death at all occupied sites and then birth at all occupied sites probabilistically, after which the population is updated for the next generation. The scheme is implemented as RF_d2S and RF_d2M following the naming conventions in Ruxton and Saravia (1998)⁴. The second scheme introduces less structure since event ordering is random and updates are fully asynchronous. We generate a list of $(site, event)$ pairs, executing these in a new random order at each generation and updating the state of system after each event. In Ruxton and Saravia (1998) this scheme is referred to as $RR1$ and is implemented here for the two forms of rate conversion as $RR1S$ and $RR1M$.

3.3.3 Rate conversion

We use two methods to convert rates: in the first, multiple births are permitted whereas in the second, an additional layer of approximation is introduced since at most one birth is performed. In the first approach, referred to as *multiple births*,⁵ we draw a number of events n_ν where $n_\nu \sim Poiss(r\Delta_t)$ where, as previously, r represents the instantaneous rate and Δ_t is the time step. This conversion makes the assumption that births within a time step are independent⁶. In the second, we assume that multiple births within a time step never occur so $Pr(n_\nu > 1) \approx 0$ and thus that $Pr(n_\nu = 1) \approx 1 - Pr(n_\nu = 0) = 1 - e^{-r\Delta_t}$ (see e.g. Fleurence and Hollenbeak, 2007). Multiple births versions of the algorithms are suffixed M and single births S .

⁴In the CA algorithm names, the first letter refers to site ordering, the second to event ordering, and the number to whether updating takes place immediately (1, because a single array is required) or with a delay (2, as two arrays are required). Specifically, R refers to the random order in which sites are visited, F_d indicates that the order of birth and death events is fixed with death occurring first, 2 indicates that two arrays are used to store the configurations to allow delayed updating, and the final letter (see 3.3.3) refers to the approach to rate conversion. The same scheme is referred to as $RF2$ in Ruxton and Saravia (1998), where no birth-first schemes are considered.

⁵This method is applied only to births as each site is only visited once per generation so death can occur at most once.

⁶A similar approach, known as τ -leaping, is used in approximations of the standard Gillespie algorithm (Gillespie, 2001).

3.3.4 General model parameters

In our experiments, we consider *Gill* as our benchmark and compare with the four CA update schemes RF_d2S , $RR1S$, RF_d2M and $RR1M$, explained in Appendix B.1. We simulate all schemes for 1000 time units on a regular square lattice of 100×100 sites using a neighbourhood consisting of the 8 nearest neighbours (Moore neighbourhood) and wrapping boundaries in the form of a torus to mimic the infinite lattice of the theoretical model⁷. We refer to the population after 1000 time units as the *final population*; in many cases this corresponds to the pseudo-steady state of the system although time to convergence depends on parameter values. We run 100 stochastic repetitions of all simulations and report summary statistics where appropriate. Except where otherwise indicated, all populations start with 1000 individuals randomly distributed across sites. In each experiment, we hold the intrinsic birth rate of organisms constant at $b = 1$ and vary the death rate in steps of 0.1 from 0.1 to 1.0 inclusive⁸, giving death-to-birth ratios, denoted δ , in the range 0.1 to 1.0. We test time step values in the set $\{2^0, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 2^{-6}\}$. The simulation code is programmed in Java and the full code release is available from <http://rebeccamancy.github.io/gillespie-cellular-automaton/>, with post-processing conducted in *R*.

3.4 Experiment 1: Single population

In Experiment 1 we explore final population sizes under *Gill* and the four CA update schemes. Final population sizes under *Gill* for the different death-to-birth ratios (δ) are shown in Figure 3.1, demonstrating small variance between runs. In general, populations with lower values of δ are larger, and for values of $\delta = 0.6$ and below, the population grows rapidly and reaches quasi-stationarity; extinction occurs slowly for $\delta = 0.7$ and rapidly for higher values (see Figure 3.2). We refer to the transition between persistent populations and population extinction that occurs between $\delta = 0.6$ and $\delta = 0.7$ as the *persistence threshold* (Adler and Nuernberger, 1994).

Comparing final population sizes and proportion of runs extinct between *Gill* and the CA algorithms, qualitative patterns differed little between the two single birth algorithms and we focus on RF_d2S . For all parameter values, RF_d2S underestimates final population size and the largest differences in mean final population size occur for the longest time steps and intermediate values of δ . The finding that the largest deviations are found for the longest

⁷Our choice of network topology is motivated by the prevalence of square lattices in the literature; we acknowledge the arguments for the use of hexagonal lattices (Birch, 2006; Birch et al., 2007; Holland et al., 2007; White and Kiester, 2008). The implemented model diverges from the theoretical model in that latter assumes an infinite random regular network rather than an orthogonal lattice.

⁸A similar approach is used by Ruxton and Saravia (1998) in choosing a death rate for their simulations; however, we test the full range of death rates throughout.

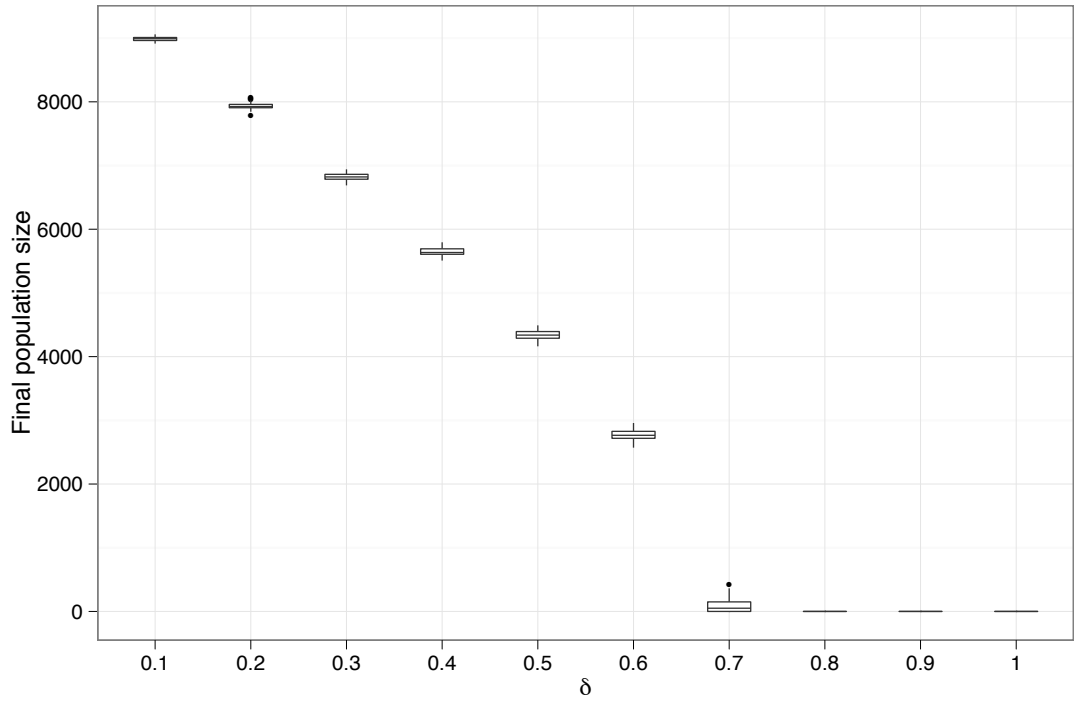


Figure 3.1: Box plot showing final population sizes for death-to-birth ratios δ under *Gill*.

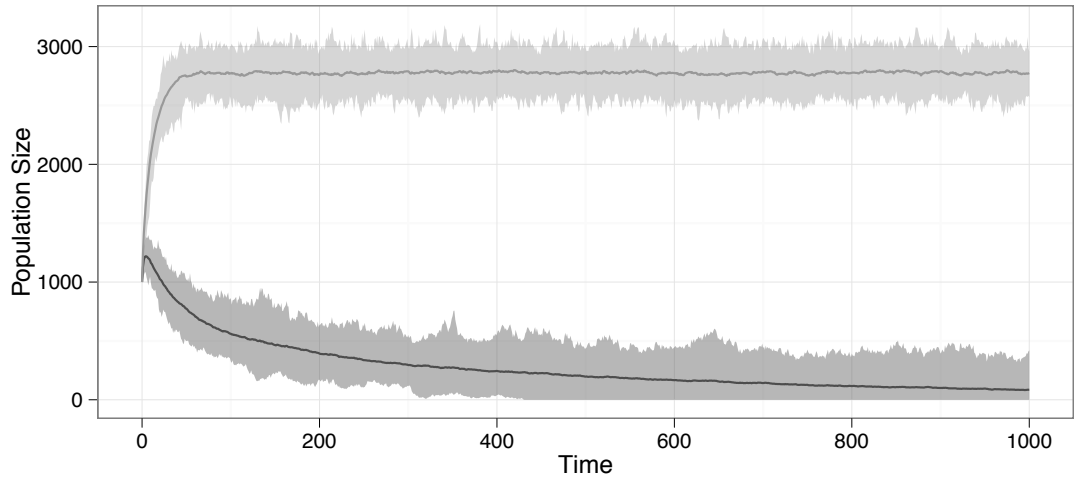
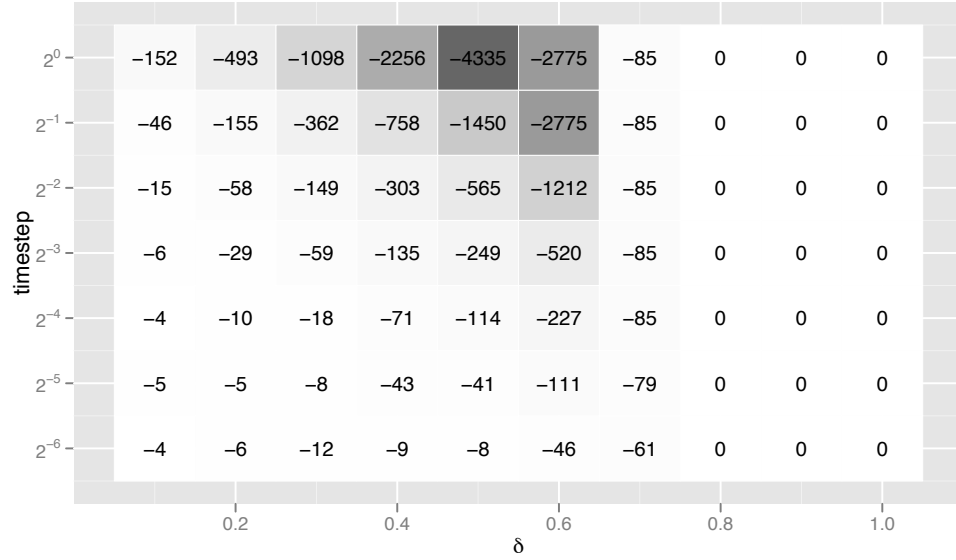


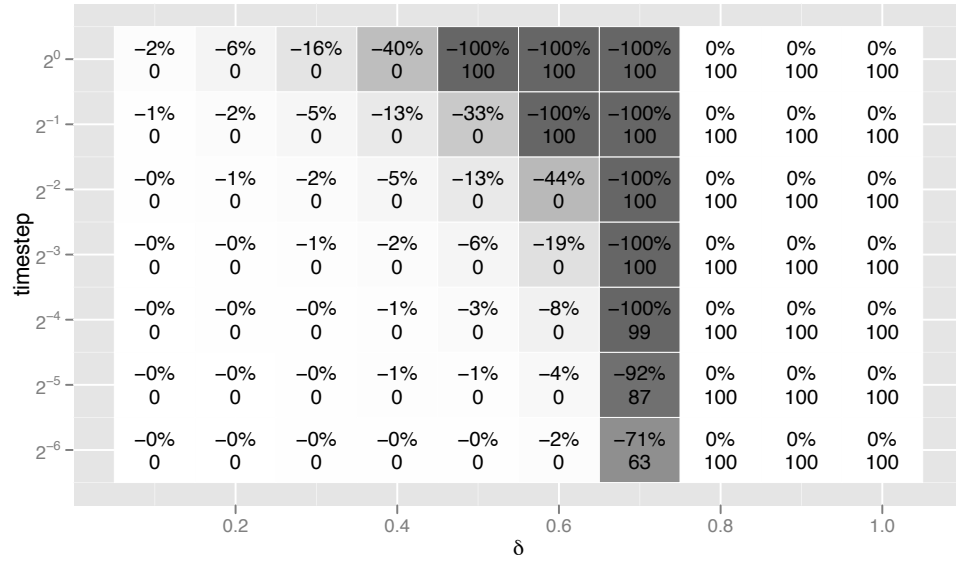
Figure 3.2: Time series plot showing mean and range for $\delta = 0.7$ (lower line) and $\delta = 0.6$ (upper line) under *Gill*.

time steps is expected. The deviation for intermediate values of δ can be explained by the occupancy level of grid: at these values, occupancy is around 50% and population size is more sensitive to differences between the algorithms than at lower δ values where lack of available sites dominates algorithm effects. Because final populations under *Gill* are small for $\delta = 0.7$, differences between *Gill* and *RF_d2S* are also small, even though all *RF_d2S*

runs actually went extinct for the four longest time steps (see Figure 3.3b for proportions of runs extinct by $t=1000$ under RF_d2S).



(a) RF_d2S : Population size underestimate



(b) RF_d2S : Algorithm bias and number of runs extinct

Figure 3.3: Plots of (a) difference in average population size (including extinct runs) for $Gill - RF_d2S$ (shading highlights difference in final population size) and (b) percentage algorithm bias (upper value in each cell; shading highlights size of algorithm bias) and number of extinct runs out of 100 under RF_d2S (lower value). Negative values represent underestimates compared with $Gill$.

Figure 3.3b shows *algorithm bias*, the proportion under-estimate of final population size compared with $Gill$ for RF_d2S . Strongest bias was found for δ close to the persistence

Table 3.1: Maximum time step size giving indistinguishable final population sizes as *Gill* (2-tailed student t-test, unequal variances at the 5% level); – indicates that final population sizes differed for all of the time steps considered; N/A indicates populations that went extinct for both simulation approaches; * indicates that this algorithm did not correctly predict extinction for these values.

δ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
RF_d2S	2^{-3}	2^{-4}	2^{-5}	2^{-6}	–	–	–	N/A	N/A	N/A
$RR1S$	2^{-6}	–	–	–	–	–	–	N/A	N/A	N/A
RF_d2M	2^{-6}	2^{-6}	2^{-6}	–	2^{-6}	2^{-5}	2^{-3}	N/A	N/A	N/A
$RR1M$	–	–	–	–	–	–	–	*	*	*

threshold, and was worst for longer time steps. The highest levels of algorithm bias are explained by the higher extinction rates of RF_d2S than *Gill*; nonetheless, even for parameters where populations did not go extinct under either algorithm, bias increases as we increase step size and towards the persistence threshold. Comparing the proportion of runs that went extinct under *Gill* (Figure 3.3b) with those of RF_d2S (Figure 3.1) shows that populations simulated under *Gill* are more resilient and this difference is most obvious at $\delta = 0.7$ (31% runs extinct under *Gill* compared with 71% under RF_d2S and 67% under $RR1S$ for the shortest time step).

Similar patterns were seen in the relationship between accuracy and parameter values for the multiple births scheme RF_d2M , except that this algorithm overestimated population sizes for all parameter values apart from $\delta = 0.7$ (Figure 3.4). In contrast, the $RR1M$ scheme gave large overestimates of final population size for all parameter values, and no run went extinct. The difference between the two multiple birth versions of the algorithms is very marked: although simulating multiple births improves estimates when updating is delayed until the end of a generation (RF_d2M) it produces large overestimates when updating takes place immediately ($RR1M$). Under the RF_d2M scheme, the increase in birth rate due to multiple births is partially compensated by the increase in death rate since all organisms are considered for death at the next time step. Under $RR1M$, births can take place into sites that have already been evaluated for death during the current generation and are therefore not evaluated for death until the following generation. In other words, the effective birth rate is increased much more than the corresponding death rate for $RR1M$; this algorithm gives a poor approximation of *Gill* and we consider it no further.

Finally, we consider the minimum time step required to generate final population sizes that are statistically indistinguishable from those of *Gill* on the basis of *t*-tests. Table 3.1 shows that RF_d2S performed slightly better than $RR1S$, but this effect is largely due to the fact that sizes under RF_d2S represent the highest point in each birth-death cycle as population

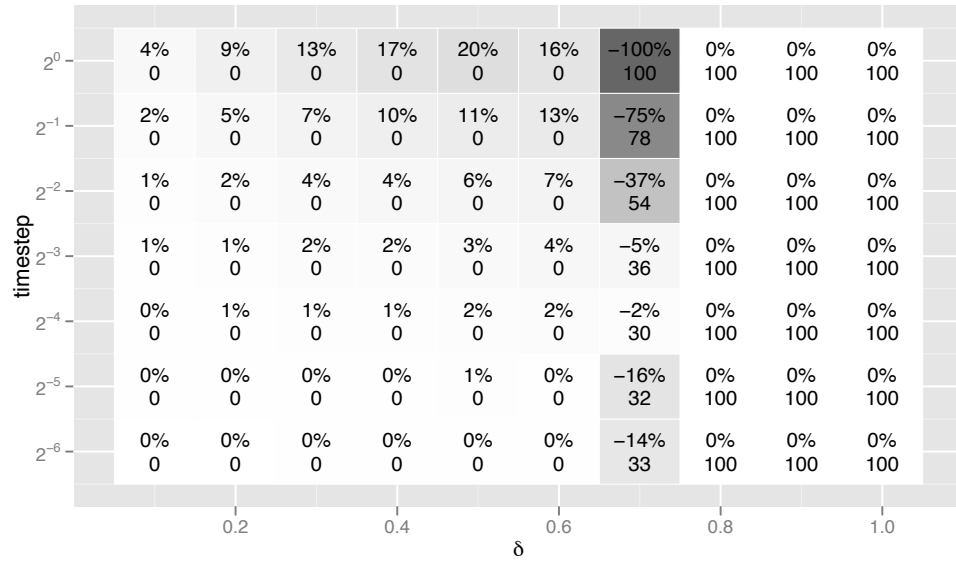


Figure 3.4: Percentage algorithm bias for RF_d2M (upper value in each cell) and number of runs extinct out of 100 (lower value). Negative values represent underestimates compared with *Gill*; shading highlights absolute algorithm bias.

size is measured after births. To simulate final population sizes that are statistically indistinguishable from *Gill*, an unidentified step size below 2^{-6} is required for *RR1S* for all $\delta > 0.1$, whereas this is the case for *RF_d2S* only for δ close to the persistence threshold. Among the multiple birth schemes, *RF_d2M* performed better (i.e. closer to *Gill*) than both single birth schemes for values of δ close to the persistence threshold, but less well than *RF_d2S* for low δ . The impact of simulating multiple births in *RF_d2M* also had a greater positive effect on algorithm bias for larger step sizes where the largest errors were found under the single births version. This is because with larger step sizes *Gill* tends to generate more cases of multiple births within the equivalent of a time step, so the single birth schemes are more inaccurate for larger step sizes.

In conclusion, Experiment 1 demonstrated relatively large discrepancies in final population sizes between the continuous and discrete time simulations, and these differed with time step and death-to-birth ratio. Time steps need to be reduced to values of 2^{-3} or smaller for final population sizes to be statistically indistinguishable from those of the Gillespie simulator, although required step size depended on both the update scheme and δ . It is therefore important for researchers who simulate in discrete time to be explicit about the time step and update scheme employed. The multiple births versions of the algorithms *RF_d2M* and *RR1M* allowed us to emulate the multiple births within a time step that can occur under *Gill*; this appeared to be a helpful strategy under *RF_d2M*, at least for values of δ near the persistence threshold, but not under *RR1M* where it introduced heavy bias. Overall, the step sizes required to accurately approximate continuous time are small, they have a complex

relationship with other model parameters and the computational cost of simulation under sufficiently small step sizes is high.

3.5 Experiment 2: Interspecific competition

We now investigate the outcomes of interspecific competition between two species. The question of interspecific competition is of importance in a range of practical contexts such as when predicting the spread of invasive species and has also been studied extensively in the mathematical biology literature (see Vandermeer and Yitbarek, 2012, for a recent example in a spatial context). For a range of simple deterministic models where competition is for a single resource, it can be shown that there are four possible biological outcomes at equilibrium: *extinction* of both species, *competitive exclusion* of species one by species two, *competitive exclusion* of species two by species one and *coexistence* (Levin, 1974). The competition model under logistic growth in the non-spatial case for species subscripted 1 and 2 can be written as

$$\begin{aligned}\dot{p}_1 &= (b_1 p_0 - d_1) p_1 \\ \dot{p}_2 &= (b_2 p_0 - d_2) p_2\end{aligned}$$

where p represents the population density (subscript zero indicating density of empty sites), b the birth rate and d the death rate. The equilibrium condition can be found by solving simultaneously for $\dot{p}_1 = \dot{p}_2 = 0$, also showing that coexistence is possible only when the two species have exactly the same death-to-birth ratio $\delta_1 = \delta_2$ and that a slight disadvantage for either species eventually leads to its competitive exclusion. We therefore also expected our spatial model to be sensitive to small deviations around this point. Since the conversion from continuous to discrete time introduces fine adjustments to birth and death rates, any differences between update schemes and methods of modelling time are likely to be apparent around this point. The decision to simulate at this point implies that differences uncovered in the experiments described below constitute a worst case scenario; however, we believe that they constitute a relevant and realistic one. For example, it is not unreasonable to assume that death-to-birth ratios of invasive species will be similar to those of native species. Furthermore, many of the studies in the theoretical literature that use CAs to validate analytic simplifications (e.g. Lion and van Baalen, 2007) are concerned with the evolution of altruism, where fine adjustments of birth and death rates for otherwise similar species are of particular interest.

We simulate interspecific competition for a range of birth-to-death ratios, holding constant the relationship between species such that $\delta_1 = \delta_2$ and with starting populations of 1000 for each species. We use the same range of birth and death rates as in Experiment 1 for

species one and allocate exactly half these rates to species two, referring to species one as having *higher population turnover*. We investigate the range of parameter values for which the CA models predict the same interspecific competition outcomes as *Gill*. In line with the stochastic nature of our simulations in which all populations would ultimately go extinct, we consider that qualitative predictions are the same when in both algorithms the same species dominates over 50% of runs at time 1000 in the sense of having larger population size.

Figure 3.5 shows final population sizes for *Gill*. Both species went extinct in all runs for values of δ of 0.8 and above. For values of 0.5 and below, the population with the higher population turnover (species one) demonstrated larger final population sizes, although the difference failed to reach statistical significance for $\delta = 0.5$. In contrast, for $\delta = 0.6$, the species with slower population turnover showed higher average population sizes. For $\delta = 0.7$, species one went extinct in most runs, and species two also tended towards extinction, but more slowly. Overall, higher population turnover is the more effective strategy for low values of δ while lower population turnover is the preferred strategy near to the persistence threshold. Time series plots showed that for persistent runs ($\delta < 0.7$), the mean population sizes were stable, although variability between runs was large.

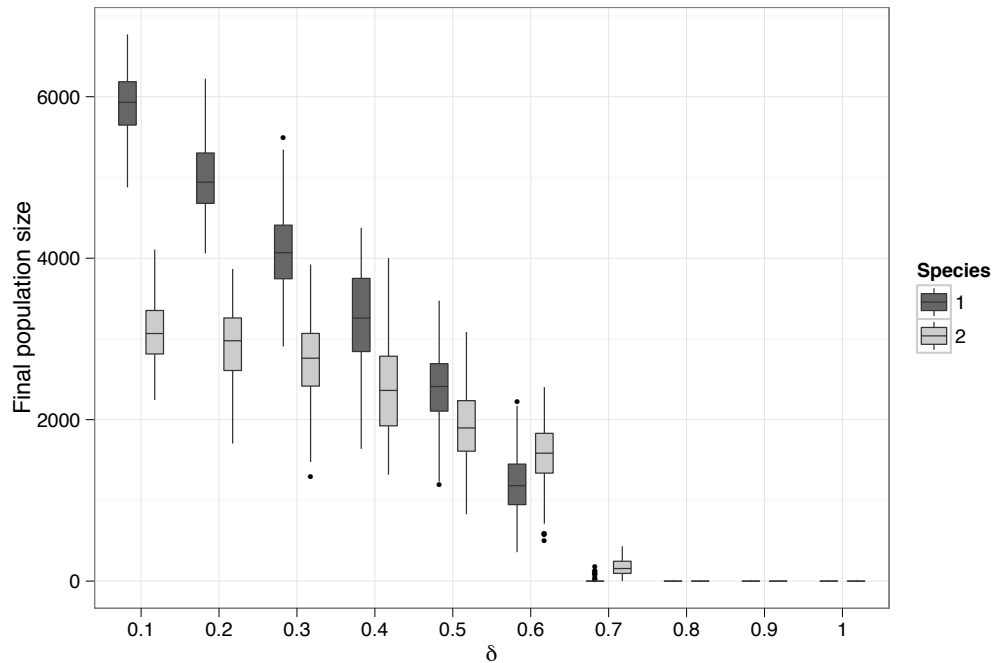


Figure 3.5: Box plot showing final population sizes for the two species under different death-to-birth ratios δ under *Gill*.

We now examine the conditions under which the CA schemes give the same qualitative predictions as *Gill*. Figure 3.6 shows the results of interspecific competitions for RF_d2S and RF_d2M (findings for $RR1S$ were very similar to RF_d2S and are not shown). Overall,

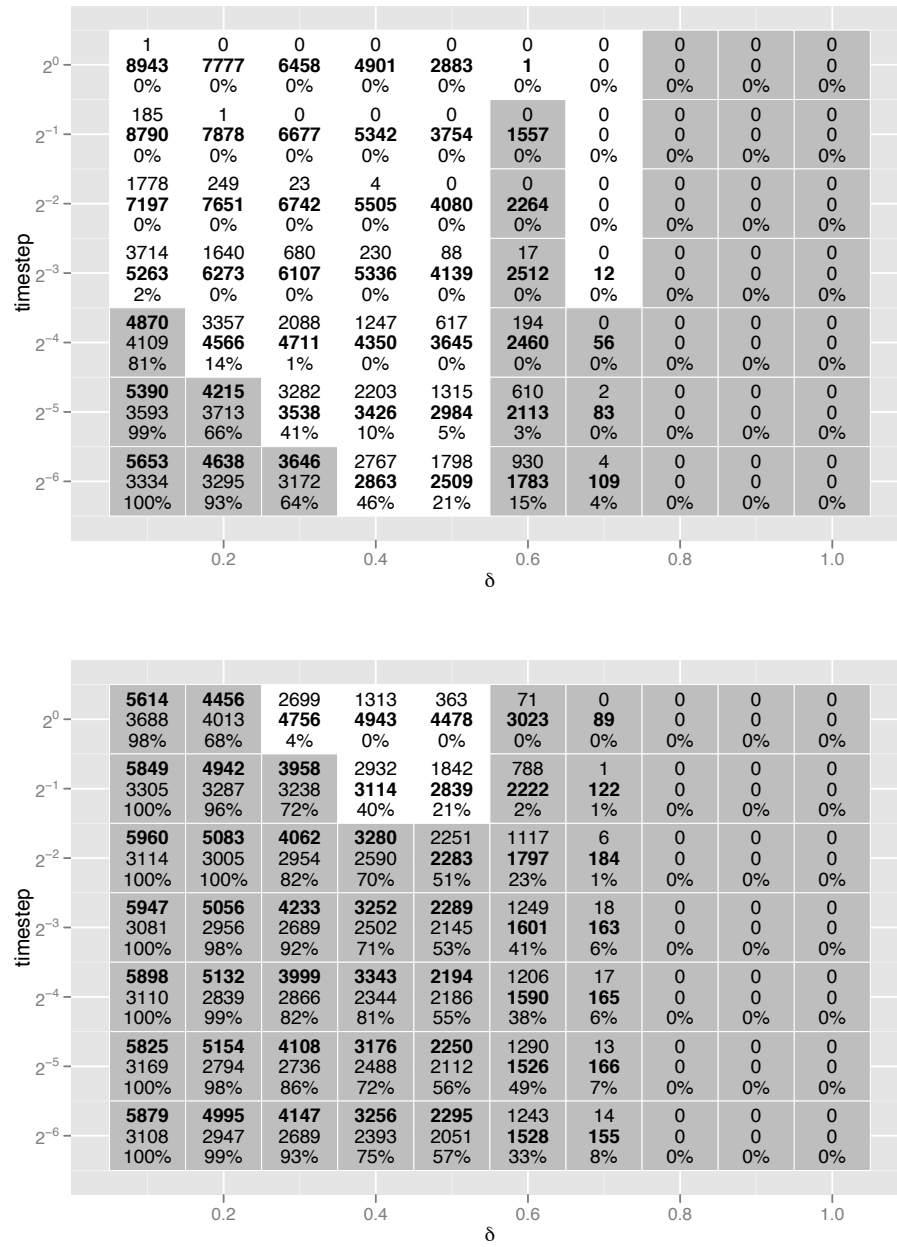


Figure 3.6: Interspecific competition for RF_d2S (top panel) and RF_d2M (bottom panel). The upper two values in each cell represent the final population size for species one and two respectively with bold type used to highlight the larger of the two values. The bottom value in each cell gives percentage of runs in which species one dominated at time 1000. Shading indicates parameter sets giving the same qualitative predictions as *Gill*.

RF_d2M performed better than RF_d2S , making the same qualitative predictions as *Gill* for a larger range of parameter values (for all values of δ for step sizes 2^{-2} and smaller). In contrast, RF_d2S predicted an advantage for the population with slower turnover for larger step sizes and all values of δ , and for lower values of δ this advantage was strong. Under RF_d2S , step size needed to be reduced to 2^{-4} before qualitative predictions concurred with

Gill for any of the parameter values for which *Gill* predicted an advantage for species 1. Figure 3.7 highlights the differences in competitive outcomes between the update schemes and shows the time series plots for *Gill*, RF_d2S and RF_d2M for $\delta = 0.1$ and for the CA models, a time step of size 1. The time series differs very considerably between *Gill* (top) and RF_d2S (middle), but much less between *Gill* and RF_d2M (bottom).

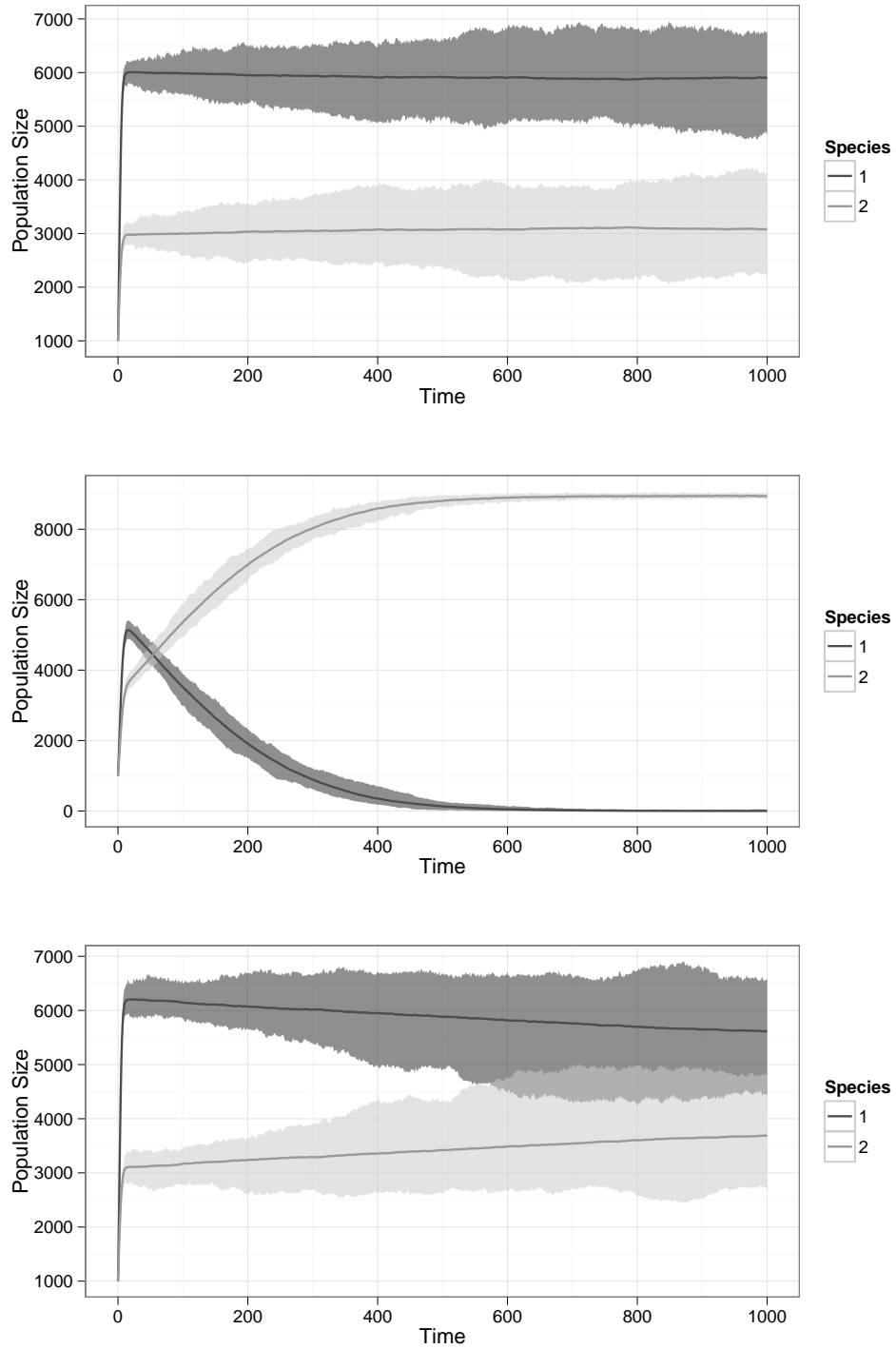


Figure 3.7: Time series plots (mean and range) for *Gill*, RF_d2S and RF_d2M , $\delta = 0.1$.

We conducted robustness testing with different initial population sizes, starting one species with a population size of 10 or 100 and the other with size 1000. Under *Gill*, the simulations showed that for large differences in initial population size, population turnover no longer dominated competitive outcomes, and populations with larger population sizes had the advantage. This effect seemed to be due to the species with larger initial size dominating the grid in early stages. In the standard *Gill* simulation with equal starting population sizes, higher population turnover led to a larger number of births at the perimeter of populated areas giving the species more rapid access to unoccupied territories where the birth rate was reduced less by competition. This effect was compounded by the resulting faster increase in population size leading to a higher species-level birth rate for the larger population. With the same starting populations, both the discrete time models tended to under-predict the final size of the population with faster turnover (species one) compared with *Gill*; these discrete time models thus performed better when species two had larger initial size since this initial imbalance in population sizes also led to dominance by species two under *Gill*.

We also conducted robustness tests to check values that deviated slightly from $\delta_1 = \delta_2$ by reducing death-to-birth ratios of species one by r_δ of 5%, 10%, 15% and 20% thus giving species one less of an advantage under *Gill*. Although the simulation paradigms differed less as we moved away from $\delta_1 = \delta_2$, differences were still apparent for *RF_d2S* for all r_δ for some time steps and values of δ . For *RF_d2M*, differences were seen for the 5% deviation from equality only. These tests showed that predictions remained sensitive to the update scheme used for death-to-birth ratios deviating from $\delta_1 = \delta_2$.

3.6 Discussion and conclusions

Our findings illustrate the importance of the decision to model time continuously or discretely, as well as the role of update scheme and step size when using CAs to approximate continuous time models or processes. Experiment 1 showed that in comparison with *Gill*, *RF_d2S* and *RR1S* underestimated final population sizes for a single species, while *RF_d2M* and *RR1M* gave overestimates. Although simulating multiple births in *RF_d2M* led to an improvement over the single births version, the same was not true of *RR1M* which made large overestimates of population sizes. Algorithm bias varied as a function of death-to-birth ratio, and differences were most apparent at values close to the persistence threshold. Experiment 2 demonstrated differences in interspecific competition as a function of update scheme, step size and death-to-birth ratio. In order to provide the same qualitative prediction as *Gill* about species dominance at time 1000, both *RF_d2S* and *RR1S* required small step sizes, although none of the values tested were sufficiently small for $\delta = 0.4$ or $\delta = 0.5$. In contrast, *RF_d2M* predicted interspecies competition outcomes fairly accurately, with a step

size of 2^{-2} being sufficiently small for all δ values.

Our findings also demonstrate the importance of timescale: it is obvious from Figure 3.7 that selecting an earlier or later time at which to sample final population sizes would lead to different comparisons between the algorithms from those shown in Figure 3.6. The dynamics differ considerably between the algorithms and while more extended runs show that *Gill* has reached a quasi-stationary equilibrium, *RF_d2S* has reached a steady state with extinction of species one in all runs, and *RF_d2M* has not yet converged. In *Gill*, species two is more sensitive to extinction due to the smaller population sizes, and in 100 longer runs (not shown), species two first went extinct around time 33,000 while no populations of species one, the species with higher population turnover, had gone extinct by time 50,000.

Although the importance in ecological modelling of selecting an appropriate CA update scheme has been highlighted previously for single species models (Ruxton and Saravia, 1998), the particular question of differences between continuous and discrete time simulations, and the role of update schemes in the approximation of continuous time processes have not been explored. This is despite the pervasive use of CA models in the literature. Our simulations demonstrate that the decision to model in continuous or discrete time matters, and that choice of CA update scheme affects both final population sizes and the outcomes of interspecific competition. As an approximation of the continuous time approach, the *RR1M* scheme failed to provide accurate predictions for any of the step sizes examined and this scheme should be avoided. More interestingly, step size had important effects on final population sizes and competitive outcomes, and to achieve accurate predictions across the range of values of δ , very small step sizes were required. Halving step size also doubles execution time (i.e. execution time scales poorly with step size, and the algorithm runs in $O(2^n)$ time where n is the inverse of the step size, i.e. $1/\Delta_t$), so selecting a step size that is sufficiently small to guarantee accurate approximation to *Gill* requires considerable computational power. We tested our code on two machines⁹, and although we did not specifically aim to optimise our code, runtime for *Gill* was shorter than all CA update schemes for time steps of 0.5 and smaller, and on the second of the two machines, was shorter than *RF_d2M* for all step sizes. We suggest that whenever continuous time is assumed the Gillespie simulator approach should be selected. As noted in the introduction, this approach has the advantage of reducing the number of modelling decisions, enhancing comparability between studies and allowing researchers to direct their focus towards biological processes rather than technical implementation. Where CAs are nonetheless employed, we recommend that researchers be explicit about both their choice of update scheme and step size, and that they conduct robustness tests to check sensitivity to these parameters.

⁹Tests were conducted by running simulations in serial with other CPU load minimised to essential processes on (1) an iMac7,1 Intel Core 2 Duo @ 2.8 GHz with cache size 4 MB running java version 1.6.0.35 and (2) an Intel(R) Xeon(R) CPU E5606 @ 2.13GHz with cache size 8192 KB running java version 1.6.0.24.

The differences between discrete and continuous time and the effect of update schemes have not been explored for stochastic simulations of interspecies competition. Our simulations demonstrated that RF_d2S and $RR1S$ required very small step sizes in order to accurately predict competitive outcomes, and for large step sizes both made large qualitative errors about species dominance for two species with the same death-to-birth ratio. In contrast, RF_d2M made relatively accurate predictions regarding species dominance with realistic step sizes. This result is perhaps surprising given the smaller step sizes required in the single species experiment and in light of Caron-Lormier et al.'s (2008) findings that in a continuous space paradigm, differences in predictions became worse for more complex systems. However, the disparity between our findings and those of Caron-Lormier et al. (2008) can probably be explained by the limitations that the space available put on the degrees of freedom in our system.

The finding that the results of interspecific competitions can be so sensitive to modelling assumptions is of concern. Although different modelling paradigms often lead to quantitatively different predictions, we also find that modelling decisions affect *qualitative* conclusions about the *direction* of competitive advantage in interspecific competition. When modelling interspecific competition, it is therefore important to choose update schemes and model parameters with care. We recommend that the decision to simulate interspecific competition in discrete or continuous time should be informed by, and ideally correspond to, the biological processes under consideration. Nonetheless, RF_d2M appears to represent a possible approximation for the continuous time lattice logistic model of population growth. The recommendations that researchers should explicitly state the update scheme and step size apply as for single species simulations.

Throughout this chapter, we have assumed that CAs are used as an approximation to continuous time models or processes, using *Gill* as a benchmark. However, if we consider both discrete and continuous time models as accurate representations of different real world systems, their predictions can also be interpreted from an ecological perspective. Our findings raise questions about the value to populations of different reproductive strategies. For example, comparing the proportion of runs that went extinct under *Gill* and the two update schemes RF_d2S and RF_d2M we find that populations of organisms with synchronised reproductive cycles are more sensitive to extinction events than those with continuous reproduction, and for long cycles (large time steps) this is true even when multiple births are possible. In the context of interspecific competition, our simulations showed advantages of slower population turnover in populations of organisms with highly synchronised reproductive cycles whereas faster turnover had the competitive advantage for continuously reproducing species, except near the persistence threshold.

In conclusion, we recommend that researchers exercise caution if using CAs to simulate continuous time processes by checking that their conclusions are not sensitive to the time

step chosen, that they justify and explicate their modelling decisions in full with reference to the biological system considered, and ideally that they use continuous time models to simulate continuous processes.

Chapter 4

The contribution of a patch to persistence in a stochastic metapopulation model

To design effective interventions for conservation biology or endemic disease control, we require accurate predictions of the effect of interventions on persistence. In simple metapopulation models, patches or populations are modelled as either occupied or unoccupied (infectious or susceptible), and interventions such as habitat creation or population vaccination can be conceptualised as patch creation or removal respectively. Several authors have proposed ways to measure the value of a patch to metapopulation persistence in deterministic and stochastic models, but the accuracy with which they can be used to predict the effect on mean time to extinction or metapopulation size in stochastic models remains untested. In this chapter, we compare predictors of patch contribution for stochastic measures of persistence from the quasi-stationary distribution. Using a numerical approach, we show that the relative effect of patch removal on mean time to extinction is well approximated by the probability of patch occupancy in the quasi-stationary distribution raised to a power $\eta \approx 2$, and provide a similar measure of the relative effect of patch removal on metapopulation size. Under the appropriate assumptions, the contribution of a patch can therefore be closely approximated by information routinely collected by ecologists and epidemiologists without the need for complex model fitting. The findings are robust to changes in the model within the spatially realistic Levins model framework, provided patch area has a similar effect on colonisation and extinction rates. The findings should help design interventions such as habitat regeneration in conservation biology or vaccination in the context of endemic disease.

4.1 Motivation

To manage biological systems effectively, either to support long-term species survival or to hasten elimination of endemic disease, we need to be able to predict the effect of potential interventions. In the case where they are carried out sequentially, it would also be valuable to be able to predict the effect of interventions on measurable outcomes in order to conduct effective adaptive management of these systems (Holling, 1978; Stankey et al., 2005).

Metapopulation models in which organisms are either present or absent from patches or populations have often been used in ecology and epidemiology. A patch can be thought of as an area of habitat or a relatively distinct host subpopulation that is susceptible to disease. The use of presence-absence models can be motivated by a lack of detailed data on population sizes or by computational limitations. In ecology, a species is modelled as either present or absent from habitat patches; in epidemiology, presence or absence of infection is considered in relation to the host populations making up the metapopulation. Interventions in real systems generally aim to optimise either persistence times or abundance: for species conservation, interventions usually aim to optimise system characteristics for long persistence times or high species abundance; for disease elimination, we design interventions with the aim of shortening time until disease extinction in the metapopulation or reducing infection prevalence.

One example of a real-world system and associated intervention is that of dog rabies described in Beyer et al. (2012). The system consists of 75 villages in Tanzania where dog rabies is endemic, and the intervention takes the form of dog vaccination programmes. The system can be conceived of as a metapopulation in which patches consist of dog populations at particular locations (dogs typically live with an owner, and owners live in villages); the size of a patch corresponds to the number of susceptible dogs in the village (e.g. similar to the landscape shown in Figure 2.4 (b)). Dogs can move between villages, and do so as a function of the distance between villages, and infectious dogs transmit rabies between village dog populations in this way. The scientific question centres on the problem of making decisions about which dogs to vaccinate, given a limited number of vaccine doses. For this purpose, it would be useful to have a guiding heuristic that enables us to prioritise dog populations for the delivery of vaccines. From a computing science perspective, the problem is interesting because finding the probability of rabies being present in each village dog population in the endemic (quasi-stationary) regime, and thus the effect of vaccinating the dogs in this village, is computationally intractable using standard techniques.

In order to design effective interventions in a metapopulation context, it would be valuable to have a heuristic to guide our selection of patches or populations to help optimise the effect of these interventions. These heuristics should help us to choose between interventions, but rely on being able to establish the contribution of individual patches to metapopulation

persistence. Obviously measures of patch contribution should therefore support accurate predictions about the importance of a patch; however, in order to have maximum impact, they should also be relatively simple to compute from available data.

4.2 Characterising patch value

In this section, we provide a brief overview of existing work to characterise the value of a patch in a metapopulation framework. To date, most of the theory has been conducted using a deterministic framework in which long-term persistence is either certain or the system ultimately tends to extinction. A common approach is to establish a persistence threshold that separates these two regimes; metapopulation size can then be evaluated in the case where persistence is guaranteed¹. Hanski and Ovaskainen (2000) introduce a measure of patch value V_i that captures the contribution of a patch to metapopulation capacity λ_M , at the threshold between persistence and extinction. Metapopulation capacity itself measures the capacity of a fragmented landscape to support long-term species persistence.

In contrast, in many standard stochastic population models, extinction is the ultimate fate of the metapopulation even if it persists for a long time. Instead of a persistence threshold, we can consider whether the metapopulation enters a quasi-stationary regime, the characteristics of that regime, and the mean time to extinction from quasi-stationarity. The quasi-stationary distribution is defined as the probability distribution of system states (i.e. particular patterns of patch occupancy) conditioned on non-extinction, once the effect of initial conditions has been lost (see Section 2.4.2 for a formal definition). Given a parameterised metapopulation model, provided the system is not too large, the full state transition matrix can be constructed and methods for finding the quasi-stationary distribution and mean time to extinction are well understood (Darroch and Seneta, 1967; Nasell, 1996, 1999, 2001a; Artalejo, 2012). Although the quasi-stationary distribution can be thought of as serving a similar role for stochastic systems as the long-run (or steady-state) solution for deterministic systems, its characteristics differ in a number of respects. As a result, it is generally unclear whether conclusions about persistence that are drawn from deterministic models apply to stochastic models, and the stochastic situation needs to be investigated separately.

The effect of particular patches on determining extinction times and the distribution of states in the quasi-stationary regime has attracted little systematic attention. Additional definitions of patch value in a stochastic model are provided in Ovaskainen and Hanski (2003b), with

¹Assuming that the deterministic system is described by a set of differential equations, a persistent system is often defined as one with a stable nontrivial equilibrium solution, in which case metapopulation size is given by a single number. A nontrivial long-run solution may take more complex forms such as cyclical patterns; however, these issues are beyond the scope of this chapter as the deterministic version of the model used here does have a stable equilibrium state.

patch value considered in relation to contributions to colonisation events w_i , to metapopulation size u_i , and to time to extinction t_i . However, there has been little exploration of how one might evaluate these measures in the absence of a fully parameterised metapopulation model, the construction of which is technically involved and requires considerable data. Furthermore, computing the full state transition matrix is also only possible for relatively small systems (for a metapopulation in which each patch has only two states and n patches, the full state transition matrix has dimension $2^n \times 2^n$), and memory limitations mean that this is infeasible for systems with more than around 20 patches. As a result, alternative approaches for estimating patch values would be highly valuable.

One option is to test the possibility of using deterministic patch value measures such as V_i to predict the contribution of patches to relevant stochastic outcomes. The extent to which V_i is a useful measure for this purpose has not been addressed. Other possible measures can be found in an applied context. Beyer et al. (2012) test a range of algorithms for vaccine allocation in a metapopulation model (with individuality), implicitly using patch value measures that combine the ‘risk’ of a patch becoming infected and of it generating new infections. A further alternative is to investigate the potential of using measures based on the quasi-stationary distribution. Day and Possingham (1995) note that for a discrete-time metapopulation model (with 8 patches both of equal and different sizes), the ranking of patches according to their probability occupancy in the quasi-stationary distribution was identical to that of the effect of their removal on 100-year extinction probability. Despite their comment that ‘this suggests that, in some circumstances, the quasi-stationary probability of occupancy is a good measure of the importance of a patch for metapopulation persistence’ (Day and Possingham, 1995, p345), the relationship appears not to have been explored further.

Our main aim in this chapter is to compare predictors of the effect of patch removal on the mean time to metapopulation extinction T_m and metapopulation size in the quasi-stationary regime S^π . These predictors are derived from both deterministic patch value measures and measures based on the quasi-stationary distribution. We evaluate their accuracy as predictors of the ‘true’ value of a patch for these outcomes, the size of the effect on measurable system outcomes, as well as the ease with which they can be computed. In the next section, we provide more information on the metapopulation model employed, the overall experimental procedures and implementation, the persistence outcome measures T_m and S^π , and the predictors tested. We then present our results and discuss the accuracy of the predictors and the relative ease with which they can be calculated.

Table 4.1: Symbols used in the SRLM

Symbol	Explanation
$p_i(t)$	Probability that patch i is occupied at time t
$C_i(t)$	Colonisation rate of patch i at time t
$E_i(t)$	Extinction rate of patch i at time t
$\mathbf{X}(t)$	Occupancy vector of the stochastic model i at time t
$X_i(t)$	Occupancy of patch i at time t in stochastic model
c, e	Species-specific colonisation and extinction parameters
A_i	Size (area) of patch i

4.3 Materials and methods

4.3.1 Model

Throughout this study we use as a case study a stochastic version of the Spatially Realistic Levins Model (SRLM) introduced by Hanski and Gyllenberg (1997). This model is illustrated in Figure 2.4 (b) and is chosen in part because it explicitly models patch heterogeneity but also because it has been used for examining questions relating to both long-run persistence and extinction times from quasi-stationarity, and because deterministic patch value measures are most developed for this model. Similar models have also been considered in relation to the quasi-stationary distribution (Day and Possingham, 1995; Pollett, 1999, 2001) and extinction times from quasi-stationarity. Although most applications have been in the area of conservation biology, it can also be used to model infectious disease (Ovaskainen and Grenfell, 2003; Viana et al., 2014). The model is relatively general, such that where findings are robust to parameter values and functional forms used, they are likely to be applicable to a wide range of practical problems.

The SRLM is a deterministic metapopulation model in which a landscape is modelled as a collection of n discrete patches that can either be occupied or unoccupied by the species of interest (notation is shown in Table 4.1). The dynamics of each patch i in the system are governed by a differential equation of the form

$$\frac{dp_i(t)}{dt} = C_i(\mathbf{p}(t))(1 - p_i(t)) - E_i(\mathbf{p}(t))p_i(t) \quad (4.1)$$

where \mathbf{p} is a vector of length n representing the probability occupancy of each patch i while C_i and E_i are functions determining colonisation and extinction rates as a function of patch occupancies at time t . We consider two versions of the model. The first is a specific version, in which the particular functional forms for colonisation and extinction rates used here are

taken from Ovaskainen (2003). Specifically, the patch colonisation rate is given by $C_i = c \sum_{j \neq i} e^{-\alpha d_{ij}} A_j p_j(t)$, where c is the species-specific colonisation parameter, $1/\alpha$ is the mean migration distance, A_j the area of a parent patch, d_{ij} is the distance between patches (and e here is the base of the natural logarithm). In other words, a daughter patch i becomes colonised at a rate C_i given by the sum of the colonisation rates c_{ij} of i by another occupied parent patch j ; these rates in turn are proportional to the area of the parent patch j and inversely proportional to its distance to the daughter patch i being colonised. The patch extinction rate is simply a function of the patch area and is given by $E_i = e/A_i$ where e is a species-specific extinction parameter, such that patch extinction rates are inversely proportional to the area of the patch². In full, this model can thus be written:

$$\frac{dp_i}{dt} = c \sum_{j \neq i} e^{-\alpha d_{ij}} A_j p_j (1 - p_i) - \frac{e}{A_i} p_i. \quad (4.2)$$

In some sections, we also consider a more general model:

$$\frac{dp_i}{dt} = c \sum_{j \neq i} e^{-\alpha d_{ij}} A_i^{\zeta_{im}} A_j^{\zeta_{em}} p_j (1 - p_i) - \frac{e}{A_i^{\zeta_{ex}}} p_i. \quad (4.3)$$

The original model can be recovered from the more general model by setting $\zeta_{im} = 0$ and $\zeta_{em} = \zeta_{ex} = 1$. The purpose of this extension can be thought of as constituting robustness testing of the findings in the earlier sections.

There are two quantities of particular interest for this study that are used in the literature on the deterministic SRLM. Hanski and Ovaskainen (2000) introduced the idea of metapopulation capacity λ_M , a measure of the propensity of a landscape to support long-term persistence of an organism. They explained that in the case where colonisation rates are given by the sum of the contributions from other patches (as in Equations 4.2 and 4.3) it can be found as the leading eigenvalue of a matrix \mathbf{M} defined as $m_{ij} = 0$ for $i = j$ and $m_{ij} = c_{ij}/E_i$ otherwise. Metapopulation capacity defines a persistence threshold $\lambda_M = \delta$ (Hanski and Ovaskainen, 2000; Ovaskainen, 2003), where $\delta = e/c$ is a species-specific parameter that captures the relative tendency of the species to colonise new patches or for patches to go extinct. The threshold gives the condition under which the differential equations describing the system have a stable non-trivial steady state. In the experiments described here, we use the relationship between λ_M and δ to normalise our results, as described below. The second quantity of interest relates directly to patch value. The patch value V_i measures the relative contribution of patch i to λ_M at the persistence threshold. This measure of patch value can be calculated directly as the difference between the metapopulation capacity of the full landscape and that of the reduced landscape $V_i = \lambda_M - \lambda_{M,i}$. Ovaskainen and Hanski (2003b) also show that

²In the ecological case, patch sizes correspond to areas; in epidemiology, they represent host population sizes.

V_i may be closely approximated by the i th element of the left leading eigenvector of matrix \mathbf{M} ; this approximate measure is called \tilde{V}_i .

We follow Barbour et al. (2014) in translating the versions of the SRLM to the stochastic context. We consider as an appropriate stochastic version the metapopulation model of Alonso and McKane (2002). This stochastic model is a continuous-time Markov chain on patch occupancy vectors \mathbf{X} described by $\mathbf{X}(t) = (X_1(t), \dots, X_n(t))$ where

$$\mathbf{X} \rightarrow \mathbf{X} + \boldsymbol{\delta}_i^n \quad \text{at rate} \quad C_i(\mathbf{X})(1 - X_i); \quad (4.4)$$

$$\mathbf{X} \rightarrow \mathbf{X} - \boldsymbol{\delta}_i^n \quad \text{at rate} \quad E_i(\mathbf{X})X_i, \quad (4.5)$$

in which $\boldsymbol{\delta}_i^n$ is the vector of length n with 1 at position i and zero elsewhere. This model was implemented in both the specific and general forms described above.

4.3.2 Experimental setup

We now explain the experimental setup and parameter values. The main aim of the computational experiment was to assess the accuracy of predictors of the contribution of a patch to the outcome measures of mean time to metapopulation extinction from the QSD T_m and population size S^π in the quasi-stationary distribution. In other words, we assume that our system starts in the quasi-stationary regime for all measures, a reasonable assumption for metapopulations that have been in existence for a long period of time. In order to assess patch contributions, we required information on the true contribution of patches, in order to compare with predictors. The true contribution of a patch was defined as the difference between T_m and S^π calculated for a baseline landscape and for the landscape obtained by removing patch i . We began by generating the set of landscapes. The true contribution, under different parameter values, was found by computing the QSD for these landscapes and calculating T_m and S^π from it; for each landscape we then removed each patch in turn and re-computed the relevant statistics. This gave us the true contribution of a patch to these outcomes. Predictors were always calculated on the basis of the original landscapes (i.e. before patch removal). Stochastic patch value predictors were calculated from the QSD, while deterministic predictors were computed from the landscape and parameter values. The overall process used is shown in Figure 4.1. We now provide more detail on the experimental procedure, initially explaining how we set up the landscapes and moving on to explain the parameter values tested.

The same set of landscapes was used for all the experiments reported in this chapter and we began by generating these landscapes. For each number of patches considered, we first generated 100 sets of patch centres located according to complete spatial randomness, within a 5×5 unit area (following Hanski and Ovaskainen, 2000). Then, for each set of patch

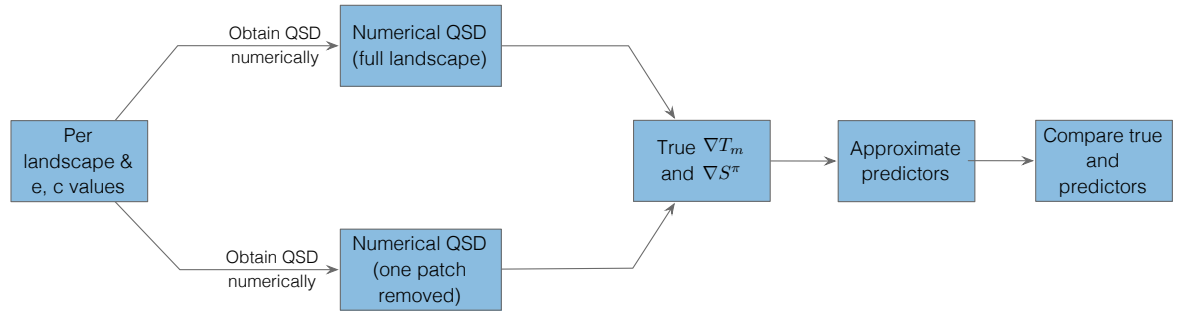


Figure 4.1: Overall experimental procedure.

centres, we generated patch areas: all areas equal, exponentially distributed and log normally distributed. Patch area for all landscapes was scaled so that the total area of each landscape was 6.25 square units, representing 25% of the 5×5 area. For each number of patches, ranging from 4 to 10, this resulted in a set of 300 landscapes (100 for each of the different patch area distributions).

The total area covered by the patches was selected as 25% in an attempt to avoid heavy overlap between patches (Ovaskainen, 2003, takes a similar approach). No additional correction for patch overlap was conducted, and although all patch centres were within the 5×5 area, patch boundaries were allowed to extend beyond it. Scaling of patch area was necessary because of the relationship between rates of colonisation and extinction and patch area, so increasing the total area with the number of patches would have increased patch colonisation rates and decreased patch extinction rates, resulting in more persistent systems for landscapes with more patches. We chose to investigate three different patch area distributions in order to verify that our results were robust to this factor (log normally distributed patch areas were used in Hanski and Ovaskainen, 2000). Before scaling, the rate parameter of the exponentially distributed patch areas was 1; for log normally distributed patch areas, the log mean was 0 and the log of the standard deviation was 1. Landscapes were generated using the statistical programming language *R* (R Core Team, 2012) and stored as comma separated value files for use in the different experiments.

The robustness of patch value predictors to changes in the overall level of persistence was considered to be important, because in the context of real interventions, we are unlikely to know the absolute level of persistence of a system *a priori*. As a result, we tested the effect of patch removal at different persistence levels as set by the ‘persistence parameters’, c and e . Because the ratio of these parameters $\delta = e/c$ sets the deterministic persistence threshold in relation to the value of λ_M for a given landscape, we decided to consider parameter combinations for e and c at this value, as well as above and below it. For the study, and for each landscape, we calculated the relevant statistics for a range of δ values, ranging from $\delta = 0.1\lambda_M$ (high persistence because $\lambda_M \gg \delta$) to $\delta = 1.9\lambda_M$ (low persistence). We tested

Table 4.2: Parameter values

Parameter	Value	Description
n	4-10	Number of patches in original landscapes
Number of landscapes	100	Number of randomly generated landscapes (per number of patch centres and patch area distribution)
Patch area distributions	3	All equal, exponentially distributed, log normally distributed; total area = 6.35 units
α	1	1/mean migration distance

this by varying both e and c ; however, the results obtained by fixing e and varying c were the same as those obtained by fixing c and varying e and we only report findings for the case of varying e while keeping $c = 1$ fixed (this has the conceptual advantage of meaning that $\delta = e/c = e$ for all experiments). In the original experiment, the migration parameter α was set to 1; in the robustness testing section, a value of $\alpha = 0.5$ was also used.

To compute the QSD, we used the standard linear algebra approach that is described in Chapter 5, where the focus is on the QSD itself is more central (see Section 2.4.2). This approach involved the construction of the full transition matrix for the Markov chain and the evaluation of relevant eigenvectors of this matrix. In order to set the parameter values e and c in relation to the metapopulation persistence threshold λ_M , the matrix \mathbf{M} was required, and this was computed and stored at this stage, along with patch values \tilde{V}_i and the steady state probabilities of the differential equations. The main code used to generate \mathbf{M} , to compute the transition matrices and find the quasi-stationary distribution was written in Matlab and run on a cluster managed by the University of Glasgow. The QSD was coded as a vector of state identifier and associated probability and stored in a comma separated value file.

Once the QSD had been computed, post-processing was conducted to calculate additional statistics and to combine the data produced in a form that was appropriate for analysis and again stored as comma separated values. Post-processing used Java for speed. The processed data were then read into the statistical programming language *R* (R Core Team, 2012), and the ggplot2 plotting system (Wickham, 2009) used to produce figures.

4.3.3 Persistence measures

Because we are interested in long-term persistence, we consider two primary outcome measures: mean time to extinction from quasistationarity T_m and metapopulation size S^π in the quasi-stationary regime. These two outcomes are chosen primarily for the practical reason

that their optimisation is often considered as an aim in interventions designed to support long-term species persistence or disease elimination. However, there are also methodological advantages to this choice. Mean time to extinction from quasi-stationarity is relatively well understood, and it is known that extinction times from quasi-stationarity are exponentially distributed, thus mean time to extinction is a sufficient statistic to characterise the full distribution of extinction times (Ovaskainen and Meerson, 2010). In addition, both can be found numerically from the transition rates matrix and knowledge of other system parameters, so stochastic simulations of system dynamics are not required.

The notion of mean time to extinction itself is straightforward. The mean time to extinction from quasi-stationarity is given by the product of the extinction rates from states with only one patch occupied and their probabilities in the QSD, summed over all one-patch states (Artalejo, 2012). It is thus calculated directly from the QSD and patch extinction rates.

The notion of metapopulation size requires more explanation as it could be assessed in a number of ways (Ovaskainen and Hanski, 2003b). We might, for example, be interested in the expected number of occupied patches in the quasi-stationary distribution or in species abundance or disease prevalence in the quasi-stationary regime. In this chapter, we consider abundance or prevalence, and population size is modelled by taking a weighted average of patch occupancy in the quasi-stationary distribution, where patch weight corresponds to the area of patch or population size. Mathematically, we can define size in the quasi-stationary regime as $S^\pi = \sum_i s_i p_i^\pi$, where p_i^π represents the probability that patch i is occupied in the quasi-stationary distribution, and s_i is a weighting of associated with the particular patch³. We choose $s_i = A_i$, so S^π represents the expected area that is occupied, or in the epidemiological context where A_i represents the number of individuals in host population i , S^π represents the expected number of cases.

Because we are interested in designing interventions, the focus of our attention is not on T_m and S^π *per se*, but rather in how these values change when patches are removed⁴. We therefore use the proportional change (or percentage reduction) in T_m and S^π , calculated by dividing the raw difference by the original value, and denote these ∇T_m and ∇S^π respectively. Thus, for a particular patch i , $\nabla T_m = (T_m - T_{m,i}) / T_m$ and $\nabla S^\pi = (S^\pi - S_i^\pi) / S^\pi$ (where $T_{m,i}$ and S_i^π denote the mean time to extinction and metapopulation size with patch i removed).

³Throughout this chapter, we use an asterisk to denote the deterministic steady state value and a superscript π to denote the value in the quasi-stationary regime. Note that by setting weights to $s_i = 1$ for all i , then S^π represents the expected number of occupied patches in the quasi-stationary distribution. The number of occupied patches might be a useful measure in a disease metapopulation situation if we need to know how many locations are likely to be infected at any one time and thus how many doctors are required to treat individuals.

⁴One might also consider the case of patch creation, but we choose to focus on patch removal, corresponding to a situation in which one needs to select the least harmful patch in the case of habitat destruction or to an intervention to remove a population from the pool of susceptible individuals through vaccination.

4.3.4 Predictors of patch value

In this section, we explain the patch value predictors that we consider. Several measures of the value of a patch have been proposed. Fairly naively, one might simply use the area of a patch (or population) A_i as a predictor of its contribution to persistence; this measure might capture patch effects relatively well when patch areas are heterogeneous and area is important in extinction and colonisation dynamics; however, for patches of equal size, it obviously provides no information. One might also choose a measure of patch centrality such as the mean distance to other patches; this may be useful when patches are clustered but would provide little information when patches are equally spaced⁵. In this chapter, we consider A_i but not patch centrality.

A possible predictor derived from deterministic models is the steady state probability of patch occupancy p_i^* ; this may provide useful information when the system has a non-zero steady state but will provide no information about patch value below the persistence threshold as it will be zero for all patches in this case. As discussed earlier, more complex measures have also been proposed for deterministic models. Also in the context of deterministic models, and as described above, another possible measure of patch value is V_i , capturing the contribution of a patch to metapopulation capacity at the persistence threshold $\lambda_M = \delta$. We use both of these measures as predictors in this chapter.

Although no measures of patch value have been proposed specifically for stochastic models of this type, Day and Possingham's (1995) finding that the probability of patch occupancy probability in the quasi-stationary distribution p_i^π is closely related to probability of extinction over a fixed period suggests that there may be a similar relationship with time to extinction. The finding in Day and Possingham (1995) related only to the ranking of patches rather than to any specific value, and the functional form of the relationship between probability occupancy and the effect on mean time to extinction is currently unknown. Further, whether and how the relationship extends to metapopulation size has not been explored. In this chapter, we consider p_i^π as a predictor, as well as various functions of it, motivated by our findings during the exploratory work carried out during the early stages of this study. Specifically, we consider $(p_i^\pi)^\eta$ for different values of η .

4.4 Findings

We are interested in our ability to predict the effect of patch removal on time to extinction and metapopulation size on the basis of what we know about a system that we assume to

⁵The situation for equally-spaced patches is not the same as that of equally-sized patches because the average distance from a patch near the edge of a landscape is higher than from a patch located near the centre.

be in the QSD. We are also interested in how likely it is that we will be able to detect the effect of patch removal. In this section, we discuss the findings of the numerical experiment and compare the different predictors, as well as the size of the effect for different parameter values. Although we tested the effects of patch removal on original landscapes with between 4 and 10 patches, comparison of the results for different numbers of patches made almost no difference to the findings, and in the sections using the specific model, we only report findings for $n = 10$. In conducting robustness testing, we chose to run the calculations on patch networks with $n = 5$ because of the number of computations required for this stage of the work.

4.4.1 Predictors of time to extinction

We begin by focusing on questions relating to the effect of patch removal on the mean time to extinction. We address two main questions: firstly, the auxiliary question of the relationship between the ratio λ_M to δ derived from the deterministic model and time to extinction in the stochastic model; and secondly, that of the relative accuracy of predictors of the effect of patch removal on time to extinction. We first consider the specific model in Equation 4.2.

In relation to the first question, as explained earlier λ_M is the persistence threshold of the deterministic model such that for $\lambda_M > \delta$, the system persists. We were interested in investigating if a similar phase transition occurred for T_m in the stochastic system as we passed through this threshold. At least for the particular model we have tested, we see no evidence of a phase transition at $\lambda_M = \delta$, and for all three patch area distributions, metapopulation extinction times grow rapidly as we decrease δ for a fixed λ_M for the landscape. This effect is strongest for uniform patch areas.

It may seem surprising that the metapopulations with the highest values of λ_M appear to have the shortest extinction times; however, this is because we set e as a function of λ_M so we are not comparing different landscapes for the same values of e and c . This finding therefore informs us that a larger landscape capacity does not fully compensate for lower persistence parameters. In other words, metapopulations with faster turnover (higher λ_M and higher δ) persist for less time than systems with slower turnover (lower λ_M and lower δ). As one would expect, we see less variability in mean extinction times for equal patch areas than for exponentially distributed and log normally distributed patch areas.

The second question is that of the best predictor of the effect of patch removal on mean metapopulation extinction time. We tested the predictors p_i^* , p_i^π , $(p_i^\pi)^\eta$, A_i , V_i for $\eta = 2$, as explained in Subsection 4.3.4. The most important finding in this section is that among the predictors tested, the effect of patch removal is best predicted by $(p_i^\pi)^2$. More precisely, the proportional change in time to extinction is well approximated by this predictor, i.e. $\nabla T_m \approx$

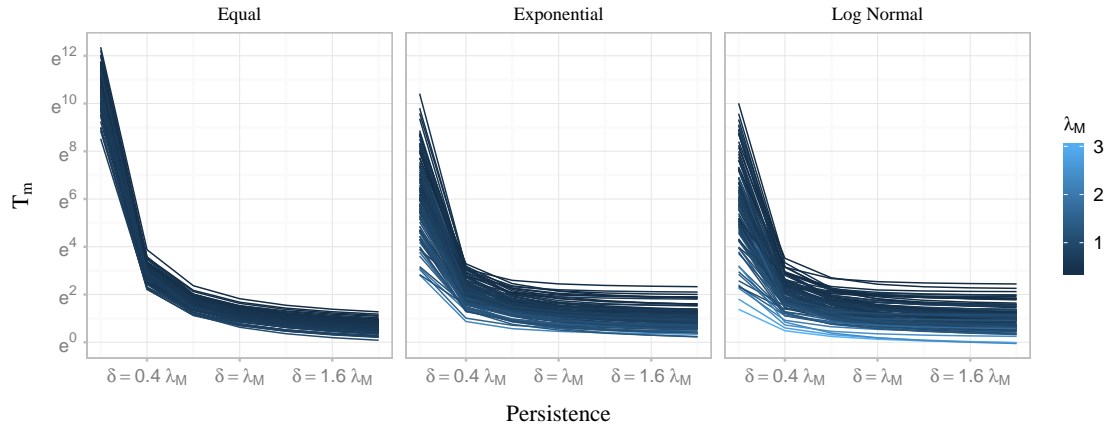


Figure 4.2: The dependence of T_m on persistence as determined by changes in values of e (note logarithmic scale; each line represents a landscape).

$(p_i^\pi)^2$. Figure 4.3 shows the relationship between this predictor and the outcome variable across persistence levels. For log normally distributed patch areas, although the correlations are lower than for equal patch areas, the relationship is roughly linear and there is limited heteroscedacity. Importantly, patches with the largest effect are predicted accurately at all levels of persistence. In the case of equal patch sizes, the approach slightly underestimates the effect of patch removal (i.e. it is conservative); however, the ranking of patches is still very good and the relationship remains roughly linear with only limited heteroscedacity.

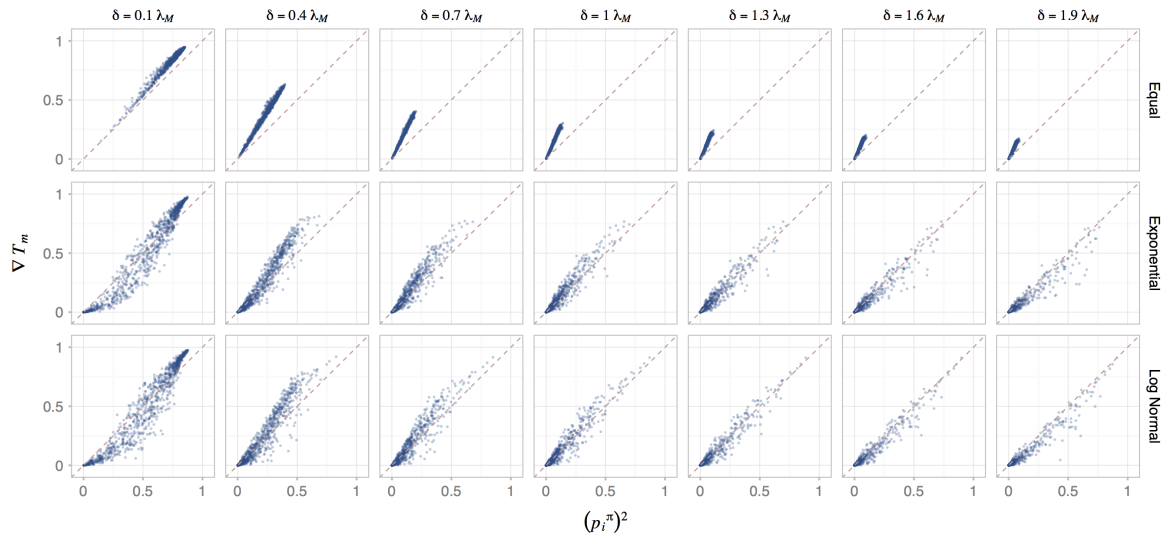


Figure 4.3: The relationship between $(p_i^\pi)^2$ and ∇T_m for $n = 10$. Parameters $\alpha = 1$ and $c = 1$ fixed; e is selected to change overall system persistence in relation to λ_M .

For comparative purposes, other predictors are shown in Appendix C.1. Perhaps the most interesting comparison is with patch values V_i . Although there is a positive correlation between

this predictor and the effect of patch removal on time to extinction, there is a lot of scatter, and in the case of uniform patch areas, the measure actually provides very little information.

In order to formalise the differences between the predictors we considered the bias and root mean squared errors of the predictions, as well as the Pearson and Spearman correlation for $(p_i^\pi)^2$ and V_i . These analyses were calculated on a per landscape basis. Whereas the Pearson correlation gives us a measure of the correspondence between the true effect of removing a patch and the raw values of $(p_i^\pi)^2$ and V_i , the Spearman correlation measures the extent to which the ordering of patches is the same⁶. Figures 4.4 and 4.5 show the bias and root mean squared error associated with the two predictors $(p_i^\pi)^2$ and V_i , demonstrating more formally the overall better predictions due to $(p_i^\pi)^2$ and reduced variance, especially for the highest levels of persistence in which there is strong bias in patch value estimates for V_i (for each landscape, there are n values, corresponding to the removal of each of the patches).

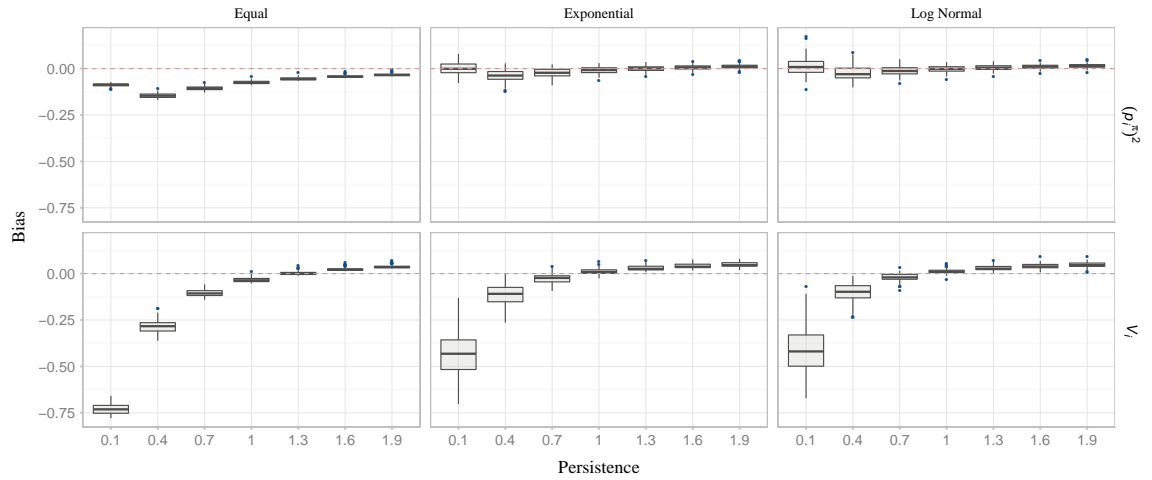


Figure 4.4: Estimator bias of $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$.

Figure 4.6 shows Pearson correlation coefficients of the two predictors V_i and $(p_i^\pi)^2$ calculated on a per landscape basis (i.e. for each landscape, a correlation coefficient is calculated for the vector containing the predictor for each patch i and the true patch contribution). Overall, it is clear that $(p_i^\pi)^2$ performs better than V_i at the level of individual landscapes, with overall better accuracy (the lowest median across all landscape types and levels of persistence was $r = 0.99$, in contrast to $r = 0.58$ for V_i) and fewer outliers. This effect is most pronounced for equal patch areas and for lowest persistence, but accuracy is also poor for V_i at the highest persistence level for all landscape types.

Figure 4.7 shows Spearman rank correlations, and shows that V_i makes better predictions about ordering than about the proportional effect of patch removal. Nonetheless, $(p_i^\pi)^2$ still

⁶This would be a useful consideration if we were interested only in prioritising patches.

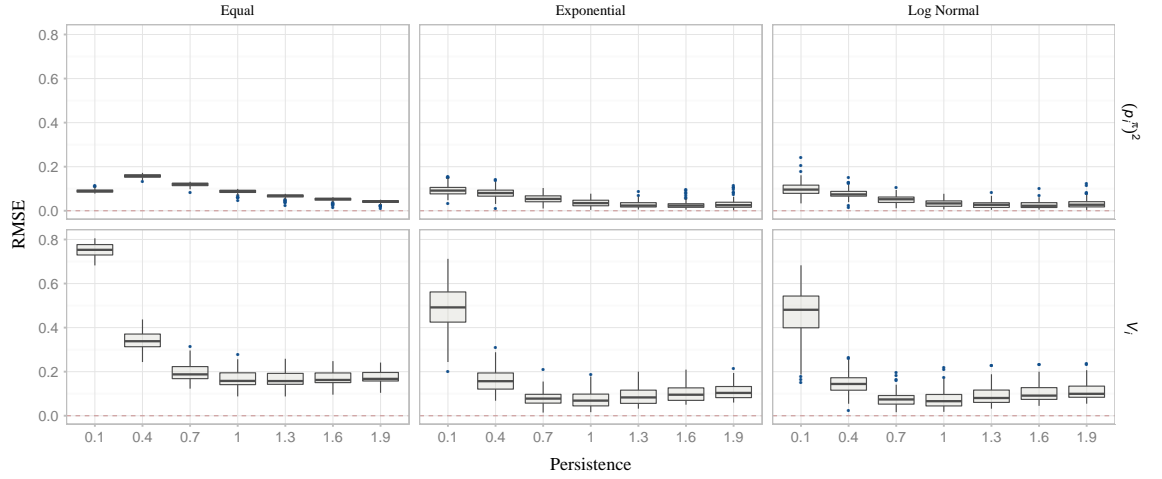


Figure 4.5: Root mean squared error of $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$.

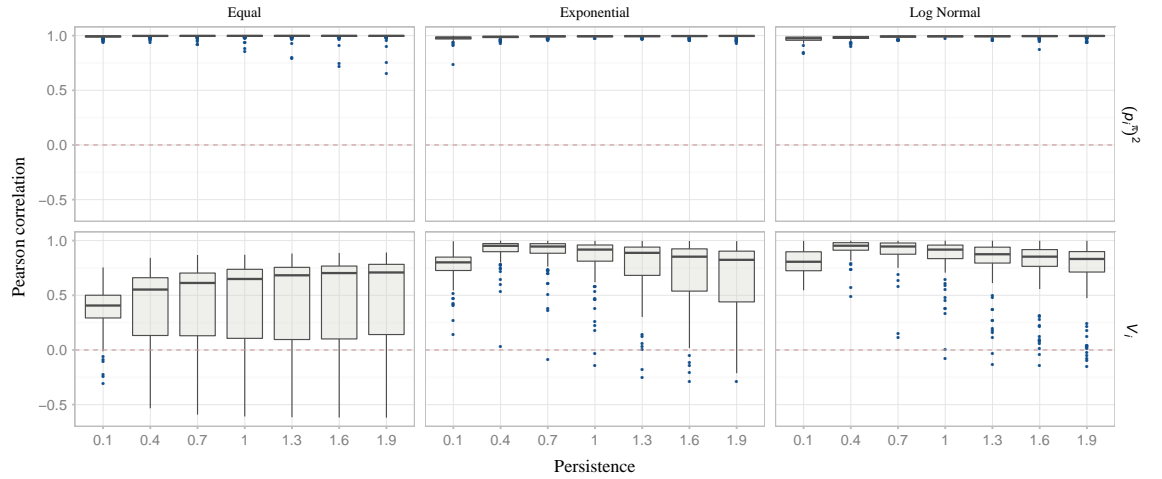


Figure 4.6: Pearson correlations for $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$.

performs better both on average and in the sense of having a lower number of outliers (the lowest median across all landscape types and levels of persistence was $r = 0.99$, in contrast to $r = 0.78$ for V_i).

For the specific model, we conclude that when the desired outcome is the expected time to metapopulation extinction, the best predictor of patch value is $(p_i^\pi)^2$. In addition to being the most accurate predictor of the effect of patch removal on extinction times and the most robust to the variations in patch area distribution and persistence, this measure has the advantage of being easy to compute from real-world data. One can compute the measure without the need to parameterise an underlying model since the value of a patch according to this measure is simply the square of the proportion of time that it is occupied in the quasi-stationary distribution. In contrast, measures such as p_i^* and V_i require parameterising the underlying

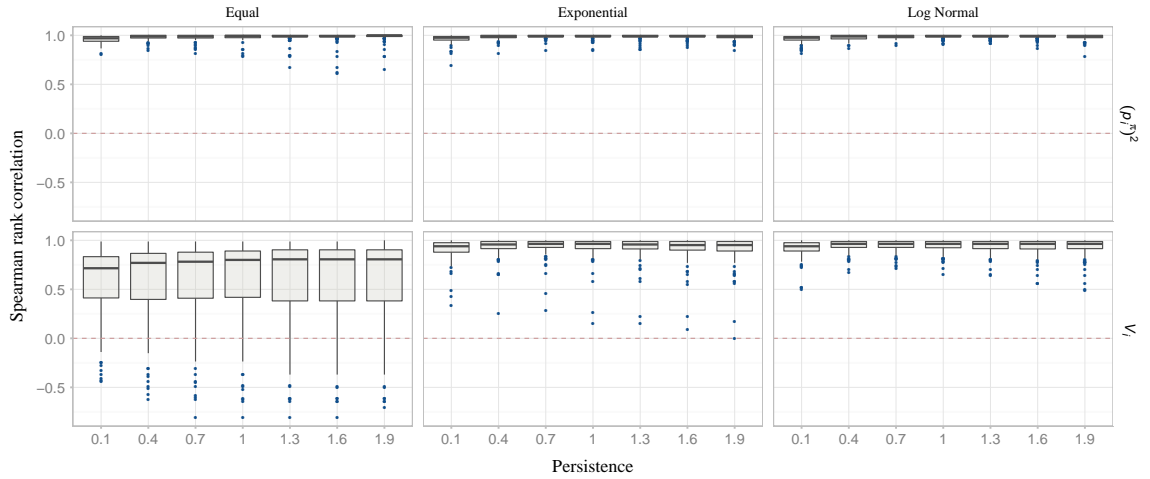


Figure 4.7: Spearman correlations for $(p_i^\pi)^2$ and V_i as predictors of ∇T_m for $n = 10$.

model, and more involved computation.

General SRLM model

To examine the robustness of the above findings, we now consider the more general model in Equation 4.3. We tested three values for each of the ζ parameters of 0, 1 and 2, and two values for the migration parameter of α of 0.5 and 1.0. Setting each of the ζ parameters to zero corresponds to eliminating the effect of patch area on this aspect of system dynamics; a value of 2 increases the importance of patch area on the relevant aspect of system dynamics. Changing the value of α changes the dependence of patch colonisation rates on distance: as α falls, the impact of distance approaches linearity; for higher values of α , the falloff in colonisation processes with distance is rapid (thus patch centrality or isolation has a stronger effect for higher values of α ; low values of α allow us to approach the mean-field approximation).

Overall, the patterns in the accuracy of the patch values were qualitatively similar for $\alpha = 1$ and $\alpha = 0.5$. Because this parameter controls the shape of the dispersal kernel (i.e. how distance from the sending patch affects the rate of patch colonisation), it is perhaps unsurprising that it is unimportant for systems with low persistence, since very few colonisation events occur before metapopulation extinction. However, even for highly persistent systems ($\delta = 0.1\lambda_M$), the qualitative patterns were similar. The measure $(p_i^\pi)^2$ generally outperformed V_i in making predictions about the effect of patch removal on mean time to extinction and $(p_i^\pi)^2$ improved as a predictor where extinction rates were more dependent on patch area. The main exception to this pattern was for parameter sets where both extinction and colonisation rates were independent of the characteristics of the patch i (i.e. $\zeta_{im} = 0$; $\zeta_{ex} = 0$). For

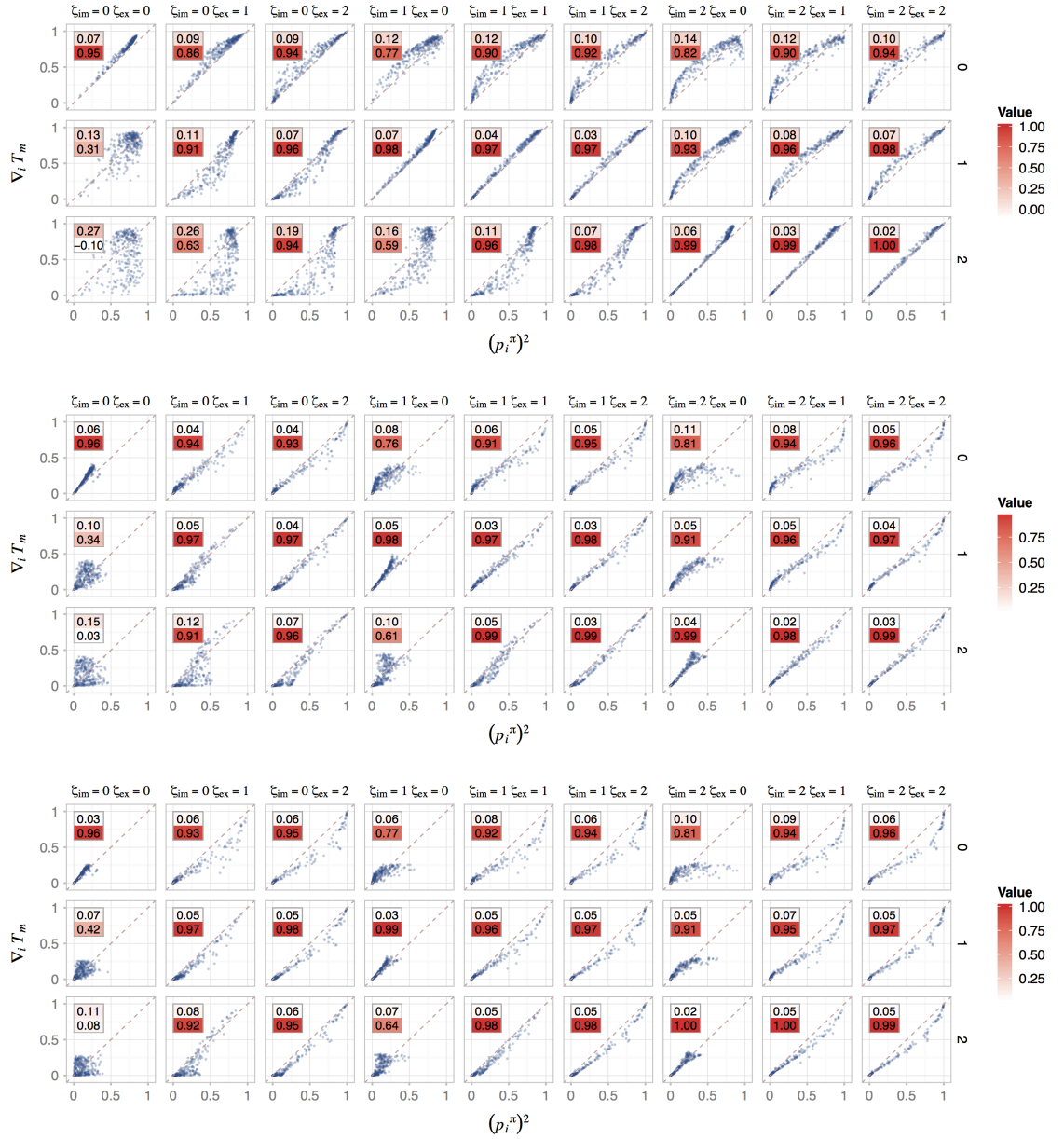


Figure 4.8: The relationship between ∇T_m and $(p_i^\pi)^2$ for different values of $\zeta_{im}, \zeta_{em}, \zeta_{ex}$ for log normally distributed patch areas. The top figure shows high persistence, the middle figure intermediate persistence $\lambda_M = \delta$ and the bottom figure shows low persistence; $n = 5$ and $\alpha = 1.0$. The coloured boxes in the top left of each scatter plot show root mean squared error (RMSE) (see top box; light shading indicates low RMSE) and Spearman correlation (bottom box; dark shading indicates high correlations). Facets labelled 0,1,2 on the right-hand side of plots give ζ_{em} values.

these parameters, only the characteristics of the sending patch were involved in determining extinction and colonisation rates. We conclude, therefore, that in the situation where one believes that patch extinction occurs at a constant rate independent of patch characteristics, and patch colonisation is determined solely by the characteristics of neighbouring patches,

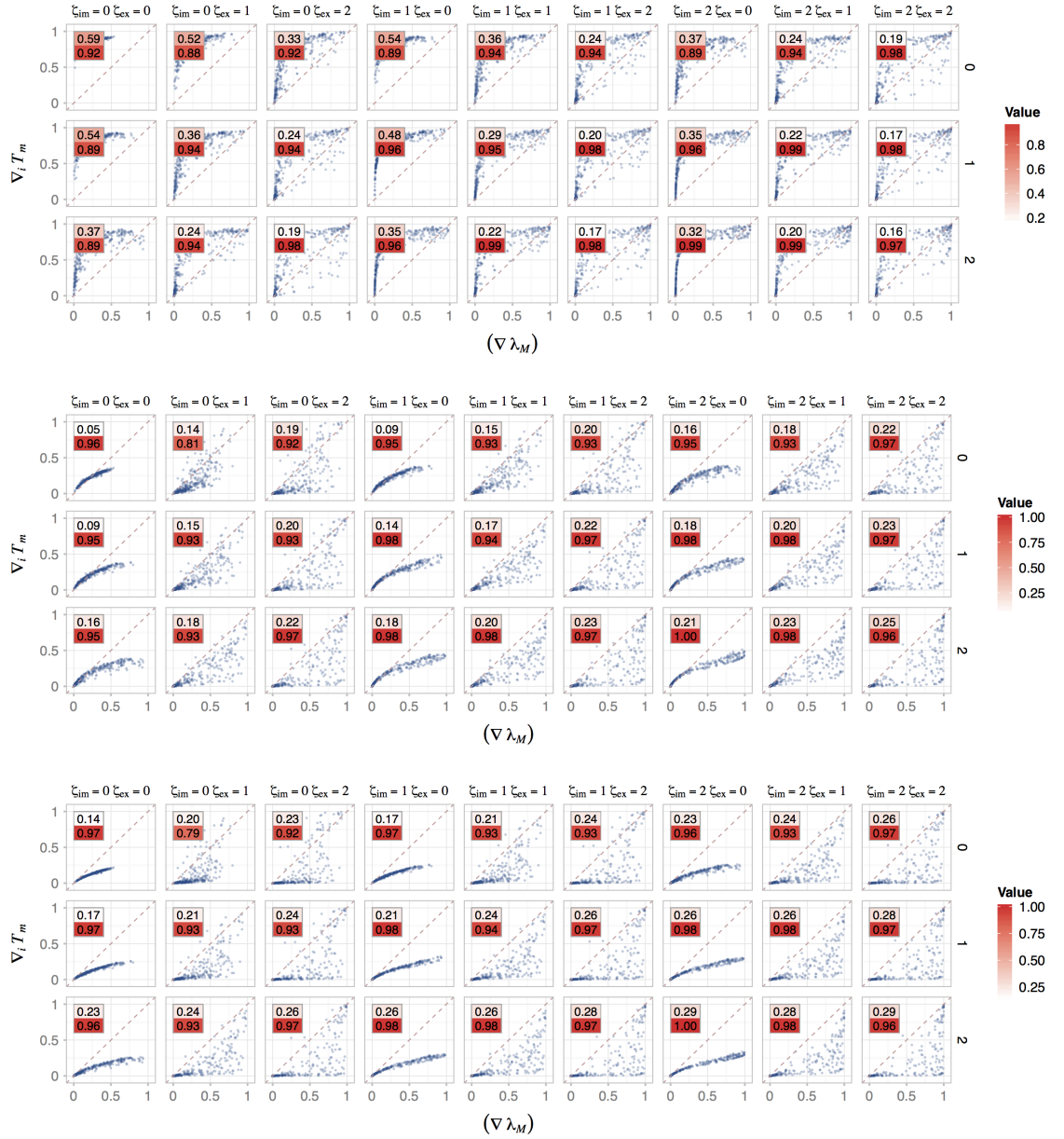


Figure 4.9: The relationship between ∇T_m and V_i for different values of $\zeta_{im}, \zeta_{em}, \zeta_{ex}$ for log normally distributed patch areas and $n = 5$ and $\alpha = 0.5$. Explanation as for Figure 4.8

the probability occupancy in the QSD should not be used as a predictor of patch value. Note, however, that these figures use the true V_i , rather than the estimate \tilde{V}_i , which will be less accurate. V_i provided better ranking of patches than p_i^π , as tested using the Spearman correlation coefficient, across a broader parameter range (compare the second of the two values in the inset boxes to the top left of the scatter plots).

Considering these robustness tests in more detail, Figure 4.8 shows the relationship between ∇T_m and $(p_i^\pi)^2$ for different values of $\zeta_{im}, \zeta_{em}, \zeta_{ex}$ and three persistence levels for log normally distributed patch areas. We can see a positive correlation between these for most values

considered. The corresponding figures for equal sized patches are omitted as these showed a strong linear relationship between $(p_i^\pi)^2$ and ∇T_m for all parameter values, with slight underestimates of patch value for all parameters and a slight trend towards better predictions at higher persistence. The comparison between the situation with log normally distributed patch areas and equal patch areas suggests that the approximation $(p_i^\pi)^2$ is robust to changes in parameter values provided patch areas are relatively uniform. The situation is more complex for more diverse patch areas. We now consider the case of log-normally distributed patch areas in more depth, noting that this distribution gives rise to highly diverse patch areas.

For log-normally distributed patch areas, a relatively clear pattern emerged in which for some values, an η of two (i.e. squaring the p_i^π values) was too high (e.g. top right of top panel), whereas for other parameters, it was too low (e.g. bottom left of top panel). That is, an exponent of 2 did not always lead to a linear relationship between patch value and ∇T_m . Further investigations demonstrated that under most of the parameter sets tested, there was nonetheless a linear relationship between $\log(p_i^\pi)$ and ∇T_m , but that the slope of the relationship was dependent on the parameters. Specifically, we noted that a relationship of the form $y = ax^\eta + \epsilon$ roughly held for all parameter sets (with different error terms ϵ) apart from $\zeta = \{0, 1, 0\}, \{0, 2, 0\}$, in which colonisation and extinction rates were dependent on patch characteristics only for other patches $A_j : j \neq i$. The linear relationship also held better where all ζ values were at least one.

4.4.2 Refined estimation of η

This suggested a way of finding an appropriate exponent η for the different parameter values ζ, α and the effect of patch removal on persistence, based on fitting a simple linear regression to find the slope of the logged data (i.e. one slope calculated for each panel). An additional multiple regression model could then be fitted in order to understand the relationship between the metapopulation parameter values⁷ and the exponent η . Should the relationship between metapopulation parameters and η prove to be linear, we would then be able to provide an estimate of patch value as $(p_i^\pi)^\eta$ where $\eta = a\zeta_{im} + b\zeta_{em} + c\zeta_{ex} + d\alpha + ep^\pi + \epsilon$, in which p^π represents system persistence (see below).

As a result of these considerations, we tested whether a better fit could be obtained by using a power η other than two, chosen as a function of the metapopulation parameters $\zeta_{im}, \zeta_{em}, \zeta_{ex}, \alpha$, and persistence p^π . Choosing a measure of persistence for this purpose is tricky: although in earlier sections we have used the relationship between λ_M and δ to char-

⁷In keeping with the earlier sections, references to (*metapopulation*) *parameters* relate to the parameters of the simulation model (i.e. the biological parameters); we use the term *coefficients* to refer to the parameters estimated using the regression model.

acterise persistence, in the real world, this information would not be available unless the stochastic metapopulation model was fitted. Since one of the valuable characteristics of the patch value measure $(p_i^\pi)^\eta$ is that the stochastic model does not need to be fitted, it would be a pity to lose that advantage in this process. As a result, we chose to use the expected proportion of occupied patches in the QSD, denoted simply p^π , as a proxy of system persistence, since this measure can also be readily calculated from raw presence-absence data. We fitted a simple linear regression to $\log(\nabla T_m)$ and $\log(p_i^\pi)$ for each panel in Figure 4.8 to obtain an estimate of the ‘true’ η value as the slope of the relationship between the logged data. We then fitted a multiple linear regression model predicting η as a linear function of the metapopulation parameters and interaction terms between these. We did this by splitting the metapopulation parameter combinations into a training and holdout set, fitting the model on the training set and testing it on the holdout set. We compared four different models, taking as the baseline model a model with a constant value of $\eta = 2$.

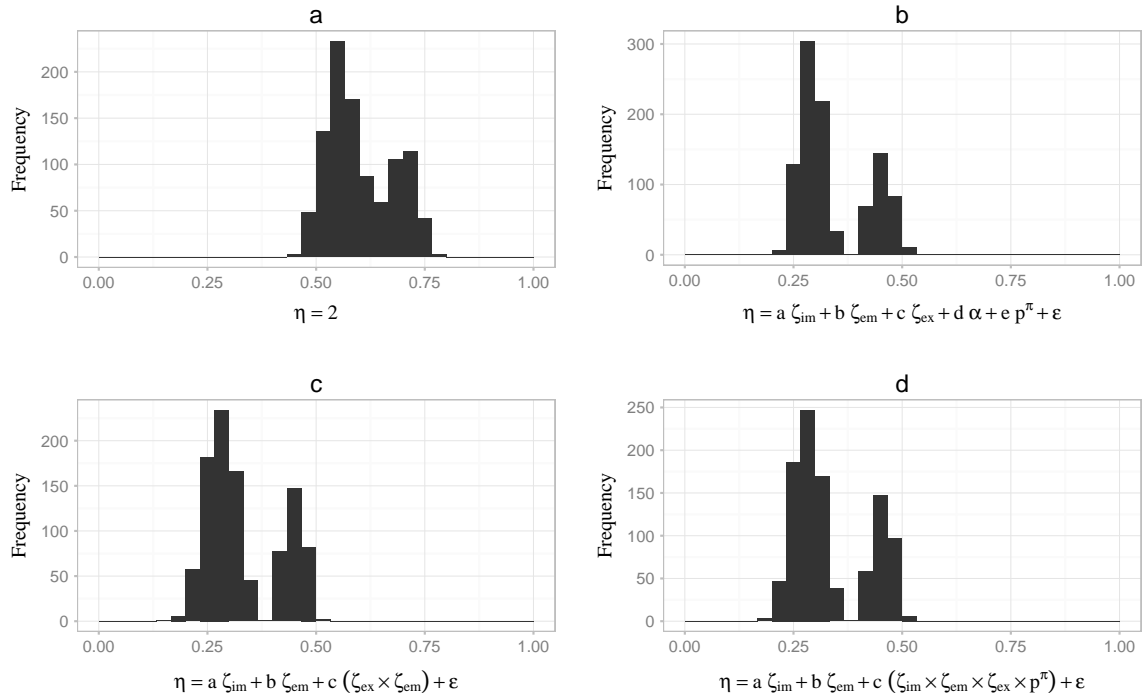


Figure 4.10: Histograms of the RMSE in predicted η values for the holdout set using (a) the baseline prediction of $\eta = 2$ and (panels b,c,d) multiple regression models fitted to the training set. The three models are fitted to the training set using a multiple linear regression of η on the metapopulation parameters, where η is found as the slope of the simple regression line of $\log(\nabla T_m)$ on $\log(p_i^\pi)$.

Figure 4.10 shows the distribution of RMSE in predictions of η on the holdout set, conducted for 1000 random segmentations of the data into training and holdout sets, and for the four different models. It is clear from the figures that including information about the parameter

values leads to better estimates of η (panels b, c, d) than simply choosing a fixed value of 2 (panel a), but there is little difference between the accuracy of other three models. All models show bimodal error distributions. This is because a small number of parameter values require very high η values (notably $\zeta = (0, 2, 1)$, especially where $\alpha = 0.5$) and in these cases, η is predicted poorly by a linear model; when these fall into the holdout set, predictions for these values are bad, skewing the mean error and generating a bimodal distribution of errors. The problematic parameter combinations represent cases where there is a specific parameter combination of ζ values, high persistence, and less of an effect of distance on colonisation (closer to complete mixing). Fundamentally, this appears to be because the relationship between η and the predictor variables is non-linear in the interaction between the ζ values, α and persistence. Fitting a non-linear model is beyond the scope of the current work but this may be a fruitful avenue for future research. In order to generate good predictions, additional data points are likely to be required at the high persistence end.

We thus conclude that using an appropriate power of p_i^π works particularly well when patches have roughly equal areas. Where this is not true, it is still effective where the effect of patch area is similar on immigration, emigration and extinction (i.e. $\zeta_{im} \approx \zeta_{em} \approx \zeta_{ex}$). However, the approach remains relatively effective provided the effects of the size of a patch are at least as strong on immigration and extinction as they are on emigration, i.e. provided that system dynamics are not dictated primarily by the occupancy status of ‘other’ patches. In the case where patch area effects are seen solely (or primarily) on emigration, the status of other patches is the main determining factor in system dynamics. As a result, $(p_i^\pi)^\eta$ fails to capture sufficient information about the system. The situation becomes increasingly bad as the system becomes more persistent. In this case, patch heterogeneity comes into play primarily when a patch is the sending patch (i.e. when a patch is playing a role in the colonisation of other patches); when the network of patches is relatively full, then $(p_i^\pi)^\eta$ overestimates the importance of a patch because the patch will rarely be required to re-colonise other patches in a well-occupied network.

The situation is particularly bad when $\zeta_{im} = \zeta_{ex} = 0$ and $\zeta_{em} > 0$, and for these parameter combinations, the relationship between $\log(p_i^\pi)$ and ∇T_m is no longer close to linear. This corresponds to a system in which patch colonisation depends only on the sending patch and patch extinction occurs at a constant rate. The first of these criteria seems biologically plausible in circumstances where establishment is not dependent on patch size (e.g. when all patches are sufficiently large to allow establishment). The second criterion of extinction rates that are independent of patch size may occur in systems where human intervention causes patch extinction at a rate that is independent of patch size (e.g. rapid ‘stamping out’ of rare disease outbreaks through effective interventions) or where patch extinction is caused by external environmental shocks that are equally effective in causing extinction in large and small patches.

One possible suggestion for those hoping to use this technique in the case of highly persistent systems where the required relationship between ζ values does not hold, might be to attempt to select the boundaries between patches in such a way that these remain roughly equal sized, although further testing is required to establish whether patch grouping in a way that it motivated by reasons that are non-biological changes the conclusions outlined here.

4.4.3 Predictors of metapopulation size

The second response variable of interest is that of mean metapopulation size in the quasi-stationary regime S^π , corresponding to the time-averaged number of infections at quasi-stationarity in the infectious disease context. Specifically, we are interested in the effect on S^π of the removal of patch i . Firstly, we examine ∇S^π as a function of persistence parameters to explore whether there is a phase transition around $\lambda_M = \delta$, and then consider a range of predictors of the effect of patch removal on metapopulation size.

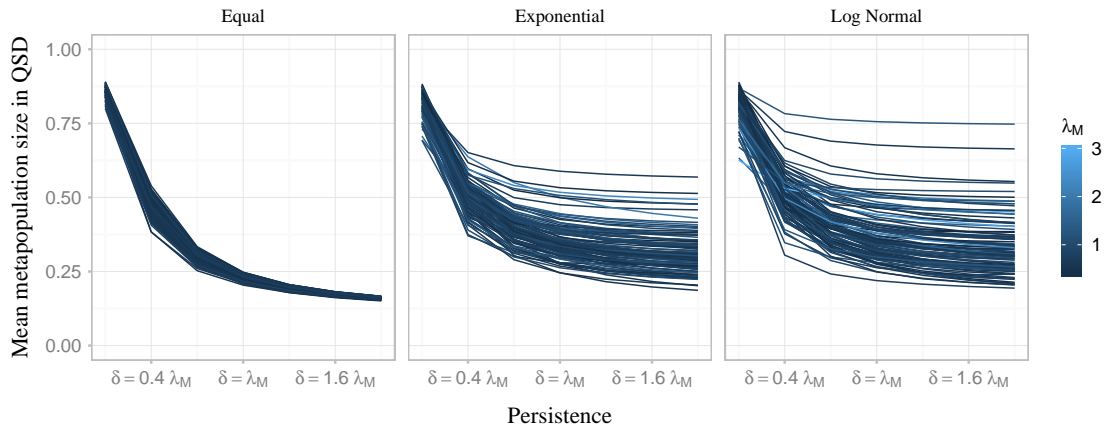


Figure 4.11: The proportion of the total area occupied in the QSD as a function of the persistence parameters.

Firstly, Figure 4.11 shows that there is no phase transition around $\lambda_M = \delta$ in the proportion of the total patch area occupied in the QSD. This result is similar to that for the mean time to metapopulation extinction.

We now consider potential patch value measures that might be used to predict the effect of patch removal on mean metapopulation size in the QSD. On the basis of the findings from the first part of the study where we predicted the effect of the removal of patch i on T_m , we hypothesised that four possible measures might be enlightening: p_i^π and the proportion of the total area in the QSD due to patch i , denoted $p^\pi A_i = \frac{p_i^\pi A_i}{\sum_j p_j^\pi A_j}$, as well as versions of both of these measures employing the square of the p_i^π values. As previously, the measures employing raw p_i^π values resulted in non-linear relationships with S^π (see Figure C.5 in

Appendix C.2). Figure 4.12 shows the relationship between the measures using the squared p_i^π values and ∇S^π .

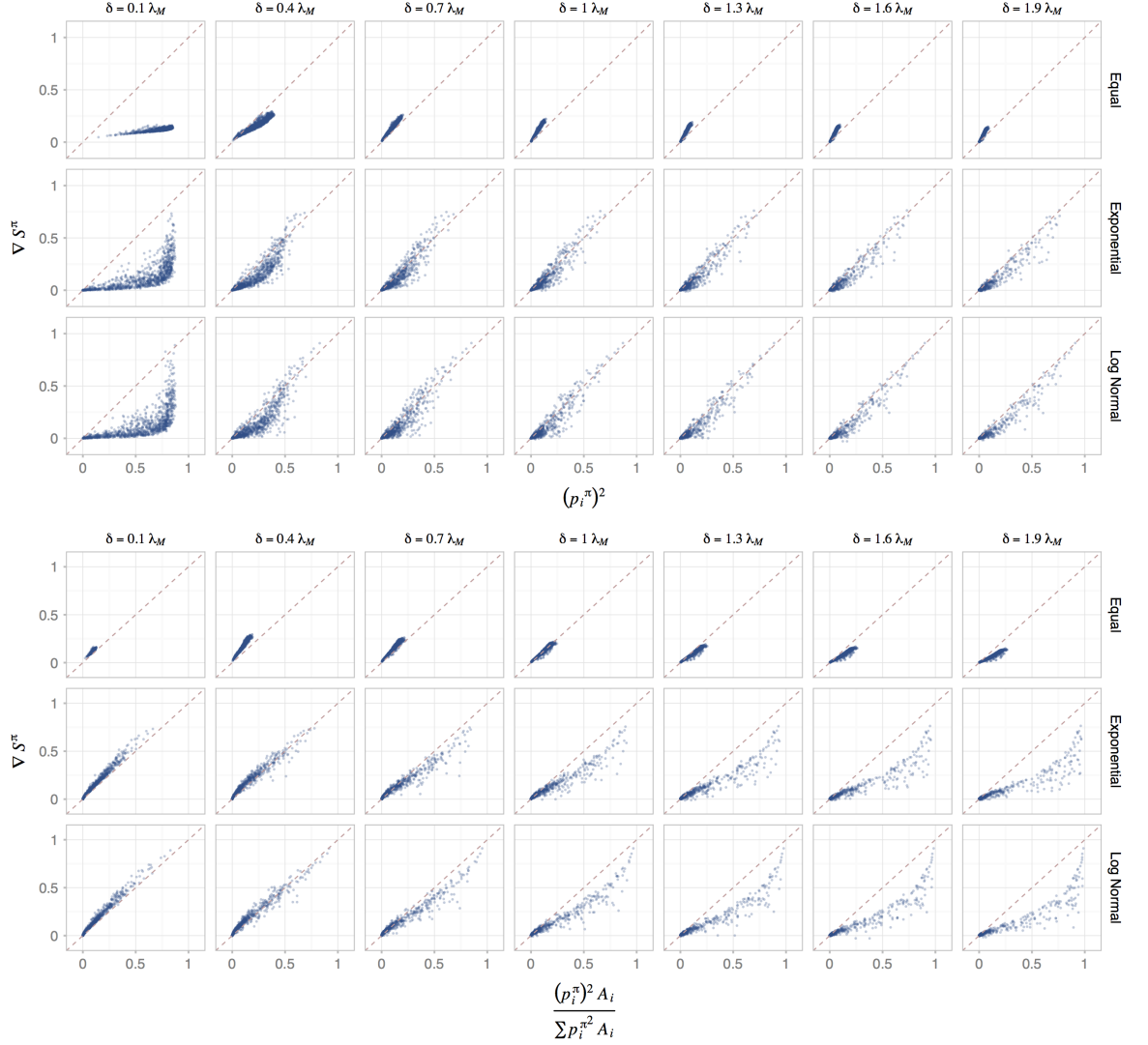


Figure 4.12: The relationship between potential measures and ∇S^π .

Although neither of the measures performed well across the range of persistence parameters, the first performed well for low persistence, and the second for high persistence. We therefore hypothesised that a combined measure might be effective. Specifically, we propose to approximate S^π by U_i , defined as follows:

$$U_i \approx \min \left\{ (p_i^\pi)^2, \frac{(p_i^\pi)^2 A_i}{\sum_j (p_j^\pi)^2 A_j} \right\}. \quad (4.6)$$

Figure 4.13 shows the relationship between U_i and ∇S^π , showing generally good correspondence. We note that this measure performs also performs better than V_i as a predictor of ∇S^π

(see Appendix C.2, Figure C.6).

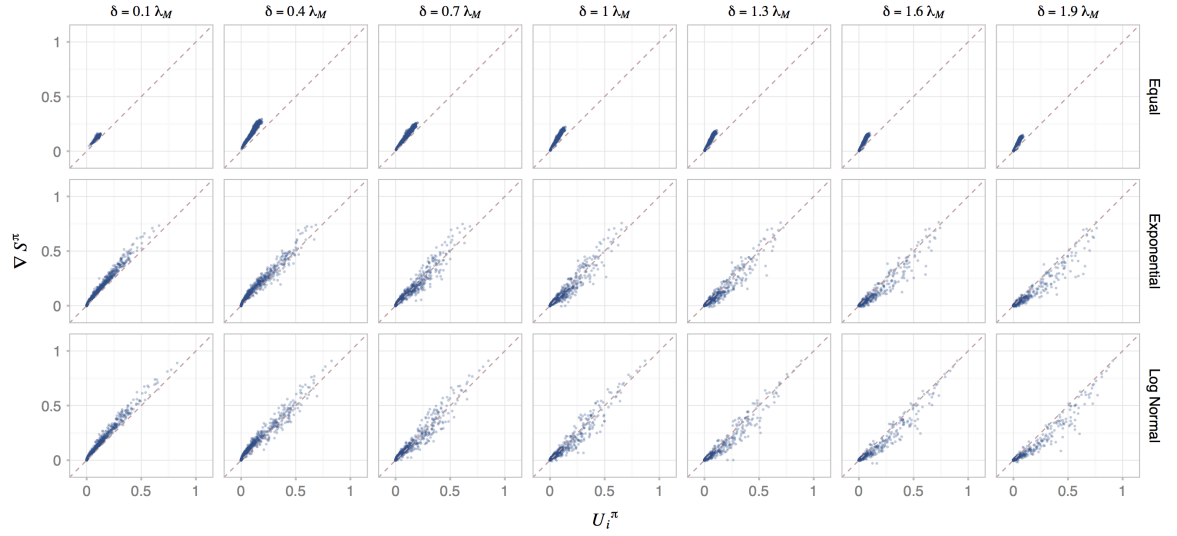


Figure 4.13: The relationship between U_i and ∇S^π .

As previously, we now show the bias, root mean squared error, Pearson and Spearman correlations for this predictor and compare with V_i . Figure 4.14 and 4.15 show the bias and root mean squared error associated with the two predictors U_i and V_i . Overall patterns in the bias as a function of the persistence parameters are very similar between the two predictors, but bias is stronger for V_i , and the direction of bias is systematically an underestimate for U_i for equal patch areas, meaning that for this situation, U_i underestimates ∇S^π . RMSE is systematically higher for V_i .

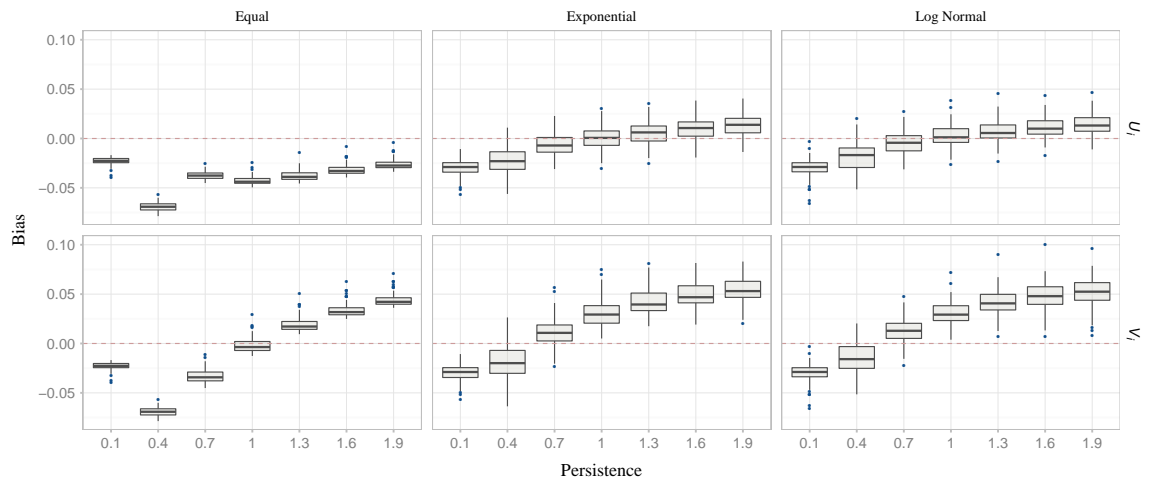


Figure 4.14: Estimator bias of U_i and V_i as predictors of ∇S^π .

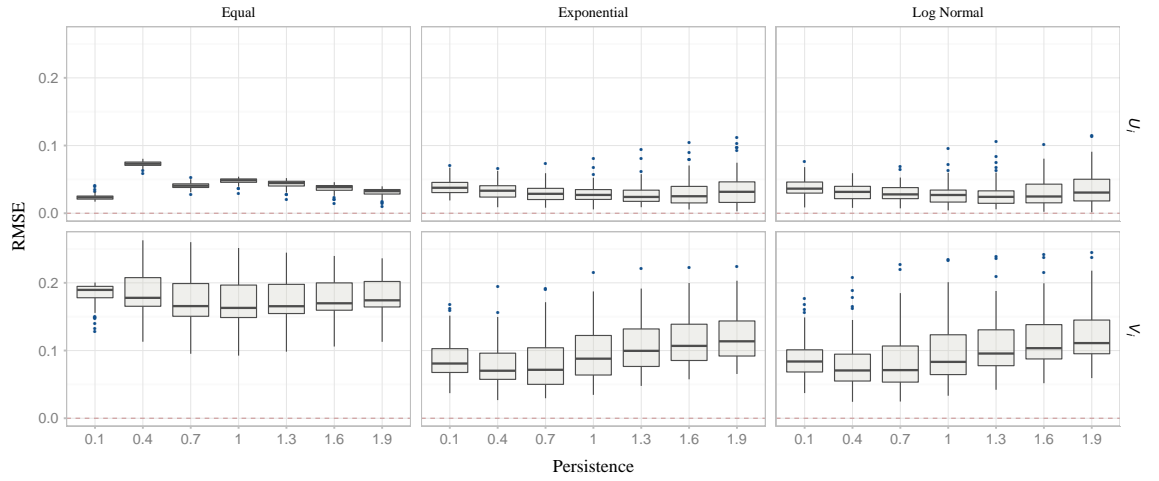


Figure 4.15: Root mean squared error of U_i and V_i as predictors of ∇S^π .

Figure 4.16 shows the range of correlation coefficients, calculated for each landscape individually, of the two predictors U_i and V_i and demonstrates that U_i performs better against this criterion across the range of persistence parameters and has fewer outliers. Figure 4.17 shows Spearman rank correlations, with similar conclusions as above, although V_i performs better in terms of patch rankings than it does for the Pearson correlations. The lower correlations for U_i for the most persistent landscapes are due to the small effect of patch removal for these values, making it difficult to detect a linear relationship. For exponentially and log normally distributed patch areas, there is a small number of outlier landscapes for which there is only a very low correlation or negative correlation (2 out of 100 landscapes for each patch size distribution).

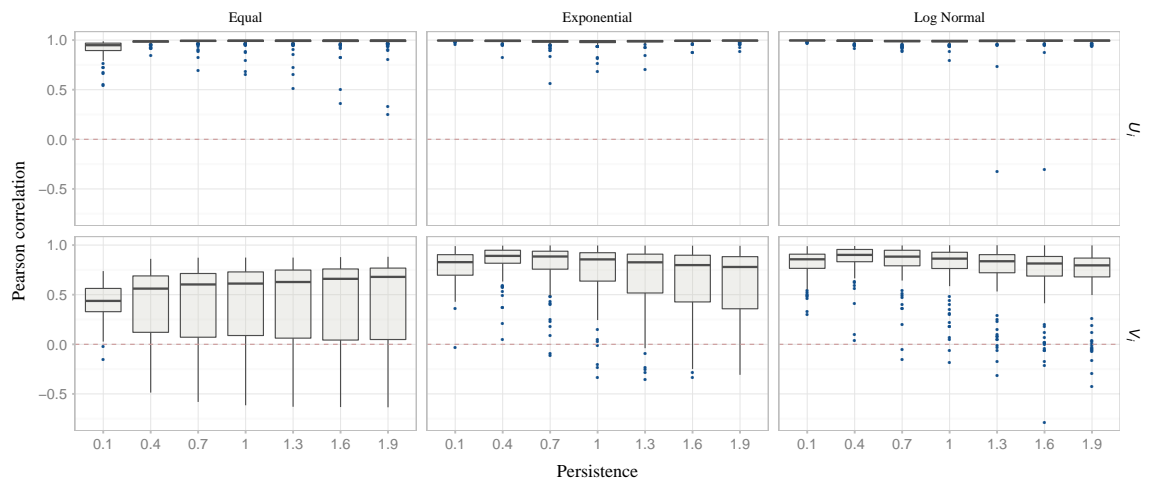


Figure 4.16: Pearson correlation between U_i and V_i as predictors of ∇S^π .

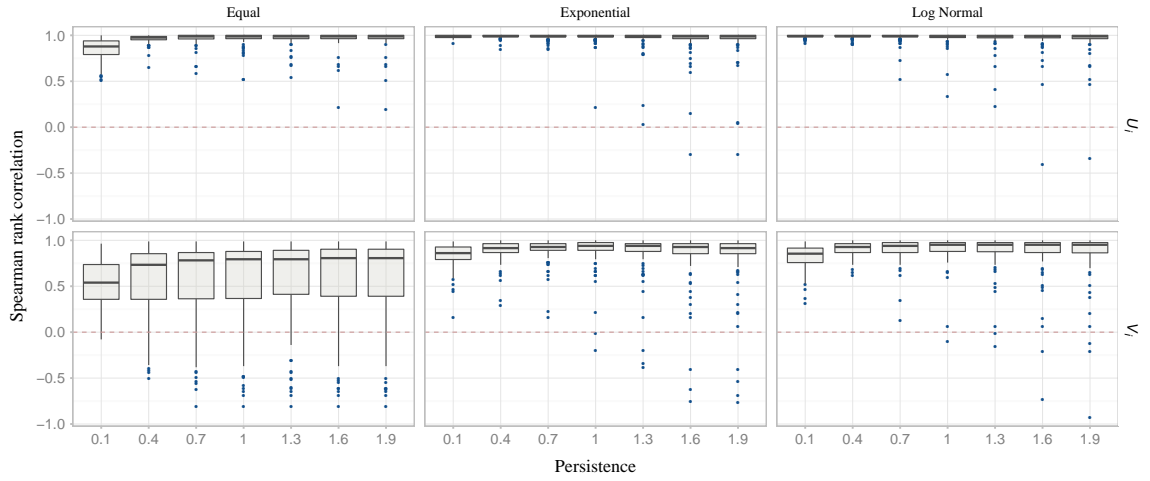


Figure 4.17: Spearman correlation between U_i and V_i as predictors of ∇S^π .

4.4.4 Predicting sensitivity to patch removal

Figure 4.18 shows the effect of patch removal on mean time to metapopulation extinction as a function of the initial occupancy levels, demonstrating that the largest proportional effect on T_m of patch removal occurs for the highest occupancy levels. This suggests that the most important effects on time to extinction of the metapopulation are likely to occur for near-full occupancy; however, it is unlikely that these changes will be detected in the short term as these systems also have the longest time to extinction, so their effectiveness cannot be used to inform incremental interventions.

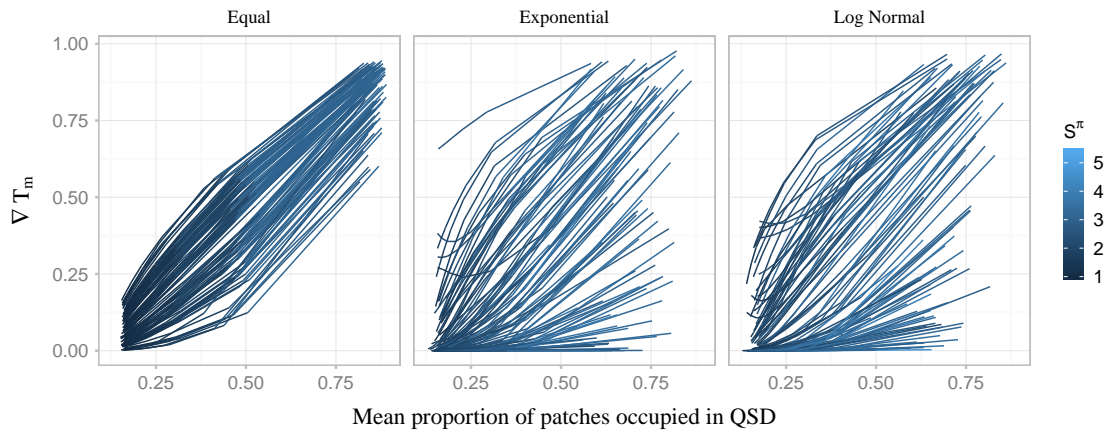


Figure 4.18: The effect of patch removal on T_m for different initial mean occupancy levels in the QSD ($n = 10$). Each line represents the removal of patch zero for one landscape.

Figure 4.19 shows the effect of patch removal on metapopulation size (prevalence or abundance) at different initial occupancy levels, demonstrating that patch removal generally had

the largest proportional effect on metapopulation size in the QSD when the mean proportion of patches occupied in the original landscape was just under half. Shading indicates the change in the proportion of patches occupied in the QSD as a proportion of those occupied in the original landscape, showing a similar trend to that of the effect on metapopulation size. This finding implies that we are most likely to observe the effect of patch removal on landscapes that are initially just under 50% occupied.

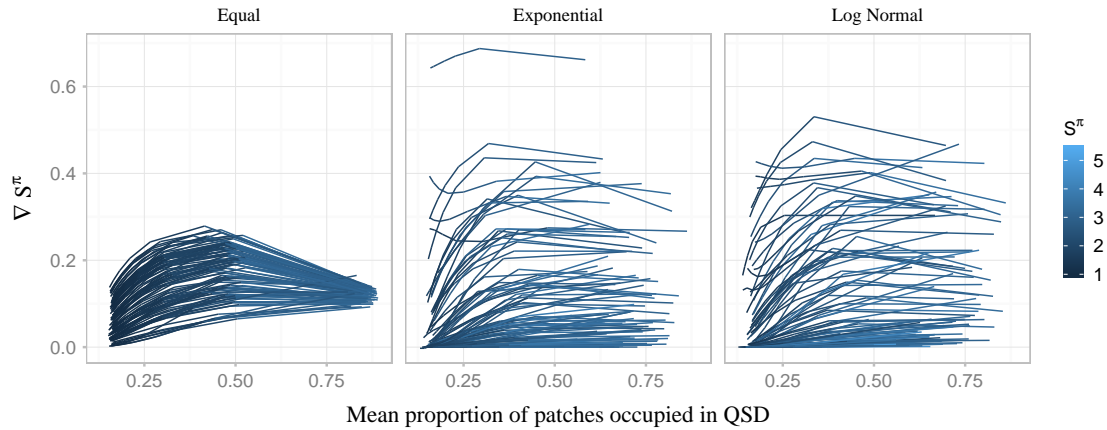


Figure 4.19: The effect of patch removal on mean metapopulation size for different initial mean occupancy levels in the QSD ($n = 10$). Each line represents the removal of patch zero for one landscape.

4.5 Discussion

In this chapter we sought to compare predictors of the effect of patch removal on the mean time to metapopulation extinction from quasi-stationarity T_m and metapopulation size in the quasi-stationary regime S^π . The findings show that the best predictor of percentage change ∇T_m is $(p_i^\pi)^\eta$ with $\eta \approx 2$, and that the best predictor of ∇S^π is our combined measure U_i . These findings strike us as surprising in their simplicity and accuracy. In effect, they tell us that if we are willing to assume only that our real world system is in the quasi-stationary regime and that the system is well-modelled by a presence-absence SRLM with functional forms for colonisation and extinction similar to those found to give good accuracy here, then the effect of patch removal on both ∇T_m and ∇S^π can be predicted with good accuracy from only information on the probability of patch occupancy.

Presence-absence data that can be used to compute occupancy probability is collected routinely by ecologists and epidemiologists. Where the above assumptions regarding quasi-stationarity and the applicability of an SRLM model with appropriate functional forms are

met, this suggests that presence-absence data can be used almost directly to choose between possible interventions, without the need for complex model fitting.

In the case where the primary object of study is a model, the findings also have practical implications. Firstly, they reduce the number of necessary calculations required for estimating ∇T_m and ∇S^π . In order to find the true value of these outcome measures, computations need to be conducted for both the original landscape and for the landscape with each patch removed, leading to $n + 1$ computations (where n is the number of patches). If one is willing to approximate patch value by the square of the probability occupancy in the QSD, a single QSD computation can be carried out. While this may not make much difference if parameter values are known fairly accurately (and the number of computations is thus $n + 1$ as above), if parameter values are estimated and one wants to conduct a sensitivity analysis, this may be a significant advantage, allowing much wider exploration of the parameter space. In addition, computing the full QSD for large systems raises other computational issues because of the number of possible states. What this finding suggests is that it may be sufficient to simulate according to an algorithm that tracks a reduced version of the QSD from which p_i^π can be accurately reconstructed, potentially reducing the required storage very considerably.

A number of possible extensions to this work would be valuable for future investigation. The most obvious extension is to conduct robustness testing for the measure of metapopulation size, along similar lines to that conducted for T_m above. Additional suggestions include testing the robustness of these findings to alternative patch groupings, systems with a larger number of patches, and whether the QSD assumption is appropriate for metapopulations that are below the critical threshold for persistence. In more detail, it would be valuable to test to what extent the p_i^π heuristics described here are robust to the way in which patches are grouped. A practical problem in applying patch models with real world data is that it is often difficult to know exactly where to situate patch boundaries, such that it would be useful to know whether this is in fact an important concern. Furthermore, grouping patches in an appropriate way would reduce the computational cost of calculating patch values in the QSD. In systems with a large number of patches (e.g. the 75 villages considered in Beyer et al., 2012), this would be highly valuable. An alternative approach in this latter case would be to test to what extent approximations to the QSD that remain plausible for the number of patches under consideration provide accurate predictions of patch value.

Another area worthy of investigation is that of the extent to which reliance on the QSD concept, when intervening on a system is valid, especially below the persistence threshold. In this chapter, we have assumed that a system is initially in the quasi-stationary regime, and that it immediately falls into the new quasi-stationary regime of the modified landscape once a patch is removed. Obviously the first assumption may not be true in the real world, where even many long-standing systems may in fact be in a transient phase; in addition, recovery of the new QSD after patch removal may be fast or slow. None of these concerns

are considered in the current work, and it is unknown to what extent their consideration may lead to qualitative or quantitative differences in the conclusions drawn. Extensions, perhaps applying the ‘ratio of expectations’ (Artalejo and Lopez-Herrero, 2010) instead of or in addition to the QSD, would therefore be valuable.

Finally, it would be helpful to analyse the predictors discussed here from a more mathematical perspective as this may help to shed light on the conditions under which the predictors fail, and also perhaps develop more accurate predictive measures (e.g. by combining p_i^π and $\sum_{i \neq j} p_j^\pi$).

4.6 Concluding remarks

The work reported in this chapter shows that even under relative minimal assumptions regarding metapopulation model structure and certain assumptions about the quasi-stationary nature of the processes under consideration, it is possible to predict with relatively high accuracy the effect of patch removal on mean time to metapopulation extinction and metapopulation size. Furthermore, the data required to enable these predictions are routinely available for ecological and epidemiological systems.

Chapter 5

Simulating quasi-stationary behaviour for heterogeneous systems

In ecology and disease ecology, we are often interested in the long run behaviour of living systems such as species distribution or endemic prevalence. In stochastic models, the notion of the quasi-stationary distribution (QSD) formalises this idea. Although the QSD can be readily obtained for small systems of arbitrary heterogeneity and approximations are available for relatively homogeneous large systems, the same is not true for intermediate sized heterogeneous systems in which states are not exchangeable. Simulating the QSD is a possible approach, but it rapidly becomes infeasible to store all states as the number of possible system states becomes large. Employing small systems in which the QSD can be found exactly, we first demonstrate the appropriateness of an existing algorithm for simulating the QSD of a particular heterogeneous system and provide a partial solution to overcome limitations of the algorithm for systems with weak persistence. We then compare a range of options for storing an approximation to the QSD that could be extended for use with larger heterogeneous systems where full storage is not possible. Although the accuracy of none of the approximations was fully robust to increases in the number of states, for specific aims, compression approaches can be selected. For examining common system states, a Bayesian clustering approach provides the best approximation among those tested. For improved robustness to increases in the number of states, and for examining time to extinction or analysis of near-extinction states, we suggest the use of a simpler system in which states are grouped by the number of occupied patches.

5.1 Introduction

In ecology and disease ecology, we are often interested in the long-run behaviour of stochastic living systems. Long-run behaviour may be interesting either in and of itself, or as the starting point for understanding future system dynamics. For example, we may be interested in long-run endemic prevalence of an infectious agent or the population size or spatial distribution of a species; alternatively, in the case where the system has already persisted for some time, we may wish to know the expected time to extinction. In classical ecological theory, long-run characteristics of a system have been investigated via analysis of the steady state of deterministic differential equations, which are also sometimes employed as a starting state for simulating future system evolution. However, in stochastic systems which may persist for a long time but must ultimately go extinct, methods for investigating long-run behaviour have been more varied.

Long-run dynamics in stochastic systems have been investigated through a variety of techniques. An intuitive approach for stochastic models is to simulate system behaviour, allowing simulations to run for a long time before measurements are taken; however, the length of the required burn-in period is often non-obvious and the question of how to account for extinct runs can be problematic. A more mathematically rigorous approach is to use the quasi-stationary distribution (QSD) of the process (Yaglom, 1947; Vere-Jones, 1969). The QSD is defined as the limiting distribution of system states, conditioned on non-extinction, as time tends to infinity. If we imagine that we simulate a large number of runs of the process, it corresponds to the distribution of states of the non-extinct runs, after the initial conditions have been ‘forgotten’. It is therefore particularly useful when the system is expected to rapidly enter the quasi-stationary regime and persist in it for a long time before eventually going extinct (Day and Possingham, 1995).

The QSD has often been invoked as having the potential to contribute to problems in ecology, disease ecology, evolution and epidemiology. For example, Pollett (1996) argues that it has the potential to be useful in wildlife management, especially for predicting persistence and the distribution of population sizes, and Ovaskainen and Hanski (2003a) point out its usefulness for understanding extinction risk. Indeed, Day and Possingham (1995) also comment that the QSD is more useful than models that only capture time to extinction as it includes additional information about the system dynamics before extinction is reached. Other authors have commented that properties of the QSD might be able to explain the stabilisation of certain biological traits among populations of some endangered species (see work by Renault, Ferrière and Porter cited in Méléard and Villemonais, 2012). In the context of epidemiology, Billings et al. (2013, p7) note that ‘the existence of a quasi-stationary distribution peaked at the endemic point produces a meta-stable state in which the population fluctuates in a neighborhood around the same endemic point’ suggesting possibilities for explaining char-

acteristics of endemic disease. In a further application in an evolutionary ecology context, Collet et al.'s (2010) work suggests applications to the stabilisation of traits under natural selection.

Despite suggestions regarding its potential contributions to applied problems, most work on the QSD is found among the mathematical literature and ecological applications are relatively rare. In fact, relatively little has changed since Pollett and Stewart (1994) commented that much of the progress at that time had been concerned with deriving conditions for the existence of quasi-stationary distributions, primarily for abstract birth-death processes.

We now describe a small number of examples of the application of the QSD to ecological problems. Among early examples, Nasell (1991) used the QSD to test whether the deterministic mosquito population size threshold for persistence of malaria in the Ross model was useful for making predictions in stochastic models, finding that the persistence threshold was unsatisfactory for small population sizes. Other early examples can be found in the literature on metapopulation processes. For example, an early application to explaining ecological patterns can be found in Gyllenberg and Silvestrov (1994). These authors investigated the potential of the QSD to explain the *core-satellite species hypothesis*, originally proposed by Hanski (1982), which predicts that at a particular point in time, a species should be either present in most sites or only in a small fraction of sites, as found in a number of natural situations. Their work showed that for a patch model with spatial heterogeneity, and under a range of parameter values, core-satellite species distributions in the QSD matched the bimodal distributions predicted by the hypothesis. In a second early application to metapopulation theory, Day and Possingham (1995) used the QSD to examine extinction probabilities of a population of malleefowl, *Leipoa ocellata*, in the Bakara region of southern Australia. These authors investigated a range of hypothetical scenarios to explore the importance of migration, the size of patches, their spatial location and their removal on persistence and patch occupancy showing that for this metapopulation, variation in patch area had a larger effect on 100-year extinction probability than the explicit spatial location of patches.

More recently, Steinsaltz and Evans (2004) have used the QSD to explain mortality plateaux, the empirically observed slowing of the death rate of individuals in extreme old age. The authors demonstrate that mortality plateaux can arise as a generic consequence of the tendency of the system to converge to a quasi-stationary distribution; they also show that early life mortality may be more flexible than generally acknowledged. In a population ecology context, Drake et al. (2011) use the notion of the QSD to predict a two-phase population extinction process according to which extinction before and after the QSD is reached differ and provide an empirical demonstration of the existence of this two-phase population extinction process with the cladoceran zooplankton *Daphnia magna*. In epidemiology, Billings et al. (2013) use the QSD to quantify the effect of intervention scenarios on disease elimination, demonstrating that for their model, elimination was accelerated more by increasing

the number of individuals treated than by treating more frequently. They also demonstrated that for their system, a random treatment time schedule led to more rapid mean extinction time than a regular schedule because as a disease nears the extinct state, rapid intervention pulses allowed by the random scheme can prevent it relaxing to the endemic regime. In our own work, explained in Chapter 4, we explored Day and Possingham's (1995) suggestion that the occupancy of patches in the QSD of a metapopulation model may be suggestive of their contribution to population persistence and found the form of this relationship.

In order to capitalise on the possibilities of the QSD for understanding biological systems, it would be highly valuable to have a method of obtaining the QSD that is appropriate for the systems under consideration. In this chapter, we focus on the technical problem of simulating the QSD for heterogeneous systems, and specifically for metapopulation models consisting of heterogeneous patches. To date, applications of QSD properties to understanding even moderately-sized heterogeneous systems have remain very limited. Among the examples described in the previous paragraphs, Gyllenberg and Silvestrov (1994) consider a metapopulation model with only three patches while Day and Possingham (1995) consider a system of eight patches and comment that the study of much larger systems is effectively impossible. Heterogeneity is either not a factor considered in the more recent examples above, or modelling is simplified to lower dimensional measures such as marginal distributions (i.e. in a metapopulation either individual patch occupancy probabilities or number of occupied patches). In fact, even Gyllenberg and Silvestrov (1994) and Day and Possingham (1995) report some of their findings in terms of the marginal distributions, rather than full system state probabilities comprised of patch occupancy configurations for the full system. Other more recent work that does incorporate heterogeneity in larger systems (e.g. McVinish and Pollett, 2013c) also presents some results in terms of patch occupancy probabilities.

A tension exists between our ability to compute and interpret the QSD for heterogeneous models and the acknowledgement that when elements of the system have different characteristics, it is insufficient to model only marginal distributions and instead the state of every patch or individual must be modelled explicitly (McVinish and Pollett, 2013b).¹ The problem of fully representing all system states in the QSD suffers from what Blanchet et al. (2014) refer to as the 'curse of dimensionality'. For example, for a landscape consisting of n patches, in which every patch can either be occupied or unoccupied, there are $2^n - 1$ possible states (excluding the extinction state). Speaking in this context, McVinish and Pollett (2013c, p694) note that, 'there is, therefore, a need for metapopulation models that are able to reflect the dynamics and heterogeneity of real metapopulations and yet are sufficiently simple to allow analysis'. Two major obstacles remain in the use of the QSD for understanding

¹Note that in the models considered in this chapter, the notion of *state* can refer to either the state of each patch (occupied or vacant) or the state of the full system. Whenever this is ambiguous, we qualify the term by referring to either patch state or system state. The terms *system configuration* and *patch occupancy configuration* refer to the overall system state.

ecological systems: its computation and its interpretation. These problems result primarily from the large number of possible system states, causing both computational difficulties and making interpretation complex.

In relation to the first difficulty of computing the QSD, the distribution is typically obtained directly from the generator matrix, a matrix describing the transition rates between states, as explained in more detail in later sections. However, since the dimension of the generator matrix grows as $m^n \times m^n$ for metapopulations in which each of n patches can exist in one of m states, evaluating the required eigenvector rapidly becomes computationally intractable. Several approximation methods have been described in the literature (see e.g. Kurtz, 1976; Pollett and Stewart, 1994; Groisman and Jonckheere, 2012; Nasell, 1991, for diffusion and truncation approximations, simulation approaches and a review of approximations). Diffusion approximations are most useful for large, homogeneous systems; the other methods may be useful for some kinds of heterogeneous systems but remain problematic for the system we consider.

In some situations with heterogeneity, the use of sparse matrix approaches may be helpful. Note that although the transition matrix for our model has dimension $m^{n-1} \times m^{n-1}$, it is sparse. Each entry of the transition matrix represents a one-step transition rate, and is set to zero if the transition is not possible; if we assume that each patch can be in one of two states (i.e. $m = 2$), then the full transition matrix has dimension $2^{n-1} \times 2^{n-1}$. In each system state, each patch is either occupied or unoccupied, and exactly one transition is possible for each patch. This means that from any system state, there are n possible transitions: exactly one of the n patches must transition from occupied to unoccupied or vice versa. Therefore, there are n possible transitions out of each of the 2^{n-1} non-extinction states, giving a total of $2^{n-1} \times n$ non-zero entries in the transition rates matrix of the chain with the extinction state removed. The generator matrix required for calculating the QSD has non-zero entries on the diagonal and thus an additional 2^{n-1} entries. To store the matrix in a sparse format, each non-zero entry is represented by a pair of integer values (r, c) indicating the row and column, along with a floating point entry indicating the rate, and thus requires $2^{n-1} \times (n + 1)$ entries.

van Doorn and Pollett (2013) review methods for the calculation of eigenvalues and eigenvectors of sparse matrices, basing their discussion on their implementation in MATLAB®. The Matlab implementation is based upon ARPACK (ARnoldi PACKAge), a numerical software library that can be used to find a small number of eigenvalues and corresponding eigenvectors of large sparse or structured matrices². To find the eigenvalues and eigenvectors of

²van Doorn and Pollett (2013) refer to LAPACK (Linear Algebra PACKage), apparently in error since this library is used for dense matrices; however, details of the methods in their discussion appear correct and correspond to those of ARPACK. Note that other interfaces exist in scientific environments such as SciPy and GNU Octave. ARPACK itself is written in Fortran77, and is available from <http://www.caam.rice.edu/software/ARPACK/>.

sparse matrices, ARPACK uses the Implicitly Restarted Arnoldi Method (IRAM; Lehoucq et al., 1998) and according to the documentation, a small number k of eigenvalues can be computed in order $O(\nu k)$ operations (where the full transition matrix has dimensions $\nu \times \nu$). A major advantage of the package is that it is able to use any matrix representation. This means that sparse matrix representations are supported, and in fact, a function can be used instead to calculate matrix entries on-the-fly, making it unnecessary to explicitly represent the matrix at all (although this would increase the number of operations).

The use of these methods significantly increases the size of the problems that can be solved. However, the QSD itself grows exponentially; as a result, the space complexity of any algorithm to compute the QSD is limited by the storage requirements of the output, and is thus at least order $O(2^n)$ for our system. Although the exact values at which these issues become intractable depends on the hardware and software used, important technical issues arise for biologically plausible problems (such as the 75 villages in Beyer et al., 2012). van Doorn and Pollett (2013) discuss approximate truncation methods for infinite systems, which may be appropriate for systems that are too large for the QSD to be found in full. These methods proceed by computing the QSD for increasing subsets of the states, where the full system is represented in the limit. However, these authors note that for systems like ours that are multidimensional, determining an appropriate state-space enumeration required to allow incremental subsetting is a non-trivial problem. In addition, pinpointing a way to conduct this subsetting of the state space that allows subsets to capture the dynamics of interest may require simulation; if a simulation is required, it seems sensible to explore methods in which the computation of the QSD is concurrent with the simulation.

Another problem is that as the number of states increases, the probability associated with any one state in the QSD typically becomes very small, and is subject to errors due to the limits of floating point representations. An approach that allows us to store counts of state visitations rather than probabilities should avoid this problem and would therefore be valuable. Finally, once a large number of states is possible, it is questionable whether any genuine biological insight can be gained from their examination as the size of the state space makes interpretation difficult. For these reasons, we now consider simulation, associated with forms of QSD compression for larger systems.

Apparently independently, Aldous et al. (1988) and de Oliveira and Dickman (2005) both describe an algorithm for simulating the QSD. The former authors prove its convergence to the QSD as time tends to infinity and the latter demonstrate that it provides accurate results for the marginal distribution of the number of occupied sites in the context of the contact process model commonly used in the physics literature. These authors thus verify the accuracy and convergence of the algorithm (in reasonable time) only for the marginal distribution of the number of occupied sites for a system with limited heterogeneity. It is therefore currently unclear whether the algorithm converges in reasonable time for simulating the full QSD. In

addition, the heterogeneity of biological systems is often much greater than that modelled by the contact process, and the convergence of the algorithm for this situation remains unverified. Furthermore, even if the algorithm is effective, although it requires less memory than the eigenvector approach, storing the full list of system states requires an array of length 2^n , and approximations are required beyond around $n = 32$. Yet real ecosystems may contain many more patches (e.g. the system considered in Beyer et al., 2012, consists of 75 villages). It would therefore be beneficial to have an approach that allows us to model the QSD for intermediate sized systems in a way that accurately reflects the true QSD, but that does not rely on storing the full list of system states.

In relation to the second difficulty of interpretation, especially in the case of heterogeneous systems, once the QSD is obtained, its dimensionality remains problematic for understanding how to use it to make inferences about real-world systems. This becomes a problem where we wish to do more than extract low-dimensional measures such as the mean time to extinction from the QSD or marginal distributions, and perhaps want to know about common system configurations or configurations close to metapopulation extinction. In cases where the distribution is highly skewed towards a very small number of common configurations in the QSD, these can be visualised individually. However, where it is not the case, it would be valuable to be able to represent the most important information from the QSD in a reduced or compressed form.

In this chapter, we explore ways to resolve the problems of computation and representation by finding a way of simulating the QSD for heterogeneous systems that employs a compressed form that is amenable to drawing biological conclusions. We conduct two main experiments. In the first experiment, we examine whether the algorithm described in Aldous et al. (1988) and de Oliveira and Dickman (2005) provides an acceptable approximation of the QSD in reasonable simulation time, not just for the number of occupied patches³ as tested by de Oliveira and Dickman (2005), but for the full QSD of patch occupancy configurations. In a second experiment, we examine a range of simplified models of the QSD that could be extended to larger systems, referred to as *compressions*, to establish whether these provide acceptable approximations. We consider the accuracy of compressions of the full QSD and on estimations of a simpler measure, the mean time extinction from quasi-stationarity, since this can be derived directly from the QSD. We use as a study system the Spatially-Realistic Levins Model (SRLM) introduced by Moilanen and Hanski (1995), a patch-occupancy model that is common in ecology and in which patches (and thus system states) are non-exchangeable.

This work should have multiple benefits for ecology and epidemiology. For example, resolving the issues of computation and representation should contribute to model parameterisation

³de Oliveira and Dickman (2005) refer to the number of occupied patches as an ‘order parameter’, using physics terminology.

by making more information accessible for fitting, to our understanding of theoretical problems such as the core-satellite species distribution, and to making predictions about endemic prevalence, all for systems in which heterogeneity and system size make standard techniques inappropriate. In the following sections, we explain the experimental procedure, present the findings of the simulation experiments, and discuss the findings in light of earlier work.

5.2 Materials and methods

The key question addressed in both experiments reported in this chapter is that of the accuracy of a simulated approximation to the exact QSD, in the context of a spatially heterogeneous system. The QSD can be found numerically from the eigenvectors of the generator matrix for systems comprising a small number of patches and we consider the QSD found in this manner as the exact QSD (it is exact up to the limits of numerical accuracy). We then compare simulated approximations to the exact QSD. The experimental protocol for the experiments is shown in Figure 5.1. In the following sections, we describe the spatially heterogeneous dynamic system used in this chapter, the numerical method used to obtain the exact QSD, the simulation algorithm, and details of the experimental setup. The description of the SRLM model and QSD are provided here for convenience; readers who have read Chapters 2 and 4 may proceed directly to Section 2.4.2.

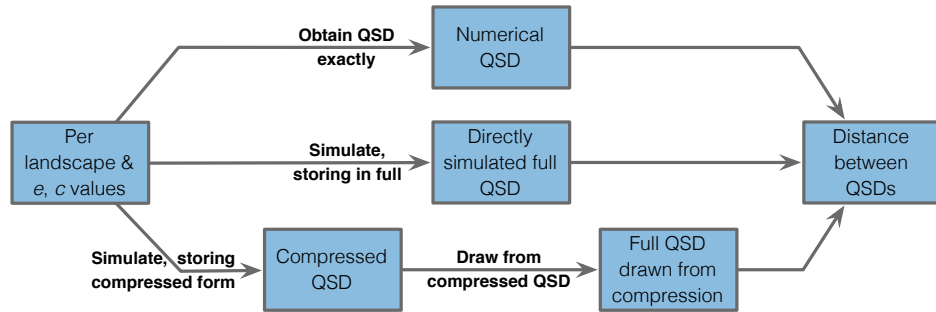


Figure 5.1: Overall experimental protocol.

5.2.1 Model

We use a stochastic, continuous time version of the SRLM as a case study throughout this chapter⁴. This model was chosen because it has been used for examining questions relating to both long-run persistence and extinction times from persistence, and because it explicitly

⁴This is the same as the less general form of the model employed in Chapter 4 and presented in Equation 4.2. Readers who have read Chapter 4 may progress to Section 5.2.2.

Table 5.1: Symbols used in the SRLM

Symbol	Explanation
$p_i(t)$	Probability that patch i is occupied at time t
$C_i(t)$	Colonisation rate of patch i at time t
$E_i(t)$	Extinction rate of patch i at time t
$\mathbf{X}(t)$	Occupancy vector of the stochastic model at time t
$X_i(t)$	Occupancy of patch i at time t in stochastic model
c, e	Species-specific colonisation and extinction parameters
A_i	Size or area of patch i

models heterogeneity in the size and distances between patches. Similar models have been considered in relation to the QSD (Day and Possingham, 1995; Pollett, 1999, 2001).

The SRLM is a deterministic metapopulation model in which a landscape is modelled as a collection of n discrete patches that can either be occupied or unoccupied by the species of interest (notation is shown in Table 4.1). The dynamics of each patch i in the system are governed by a differential equation of the form

$$\frac{dp_i(t)}{dt} = C_i(\mathbf{p}(t))(1 - p_i(t)) - E_i(\mathbf{p}(t))p_i(t) \quad (5.1)$$

where \mathbf{p} is a vector of length n representing the probability occupancy of each patch i while C_i and E_i are functions determining colonisation and extinction rates as a function of patch occupancies at time t .

The particular functional forms for colonisation and extinction rates used here are taken from Ovaskainen (2003). The patch colonisation rate is given by $C_i = c \sum_{j \neq i} e^{-\alpha d_{ij}} A_j p_j(t)$, where c is the species-specific colonisation parameter, $1/\alpha$ is the mean migration distance, A_j the area of a parent patch, d_{ij} the distance between patches (e here is just the natural logarithm of 1). In other words, a daughter patch i becomes colonised at a rate C_i given by the sum of the colonisation rates c_{ij} of i by another occupied parent patch j ; these rates in turn are proportional to the area of the parent patch j and inversely proportional to its distance to the daughter patch i being colonised. The patch extinction rate is simply a function of the patch area and is given by $E_i = e/A_i$ where e is a species-specific extinction parameter, thus patch extinction rates are inversely proportional to their area.

5.2.2 Simulation algorithm

In this study, we compare the exact QSD, computed using the standard linear algebra approaches explained in 2.4.2, with simulated approximations. Several algorithms have been proposed for simulating the QSD (see de Oliveira and Dickman, 2005, for a review). We use the simulation algorithm described in Aldous et al. (1988) and de Oliveira and Dickman (2005) to simulate the QSD. According to the algorithm, the Markov chain X_t is approximated by a chain X_t^* . First, an initial state is drawn uniform randomly among the $2^n - 1$ permissible states. The simulation of X_t^* then proceeds as would be the case for X_t until the chain enters the absorbing state (metapopulation extinction); X_t^* then transitions to a non-absorbing state (as described below) and continues in the usual manner with the same transition probabilities until it again visits the absorbing state. At regular census intervals, the state of the chain X_t^* is recorded. Storage is conducted in different ways in the compressions (see Section 5.2.3), but in the full storage variant, this leads to a histogram of state counts representing the current approximation to the QSD. When X_t^* enters the absorbing state, a new state is chosen by drawing from the distribution of state counts in the current representation of the QSD (although see below for a slight adjustment used in early stages of simulation). In order to progress between states, we employ a continuous-time spatial version of the Gillespie algorithm (Gillespie, 1977), akin to that described in Mancy et al. (2013) and in Chapter 3. Note that although the simulation itself is in continuous time, states are stored at discrete census points. This approach facilitates comparisons with approximations described later. The algorithm is shown in pseudo-code in Algorithm 5.1.

We implemented two versions of the full storage algorithm for the purposes of code verification. In the first, using Matlab, we constructed the full $2^n - 1 \times 2^n - 1$ transition rates matrix for the Markov chain X_t^* , drawing rates directly from the matrix to simulate state transitions. In the second version, coded in Java, we calculated patch colonisation and extinction rates on the fly on the basis of current patch occupancy. The latter version increased the number of calculations required during the simulation because rates needed to be recomputed after each event (rather than being computed once and stored in the transition rates matrix); however, it reduced the memory overheads since only n transition rates needed to be stored. The storage requirements of the first version scaled poorly and the latter version (the computational complexity of which is considered in more detail in the following section) scaled significantly better as the number of patches increased so this version was used in the experiment. The process of coding two versions, using two different languages (and their associated procedural / scripting versus object-oriented paradigms) and different data structures, led to greater confidence in the correctness of the code as differences in output led to more thorough checking of both versions.

Algorithm 5.1 *simQSD*. The simulation runs until the specified number of events is reached. For each event, a time lapse is simulated and the previous state recorded in the QSD representation according to the number of census intervals that have passed while in this state, then the current time is updated. A new state is simulated and the system is updated accordingly unless this represents the extinction state, in which case a new state is drawn randomly from the non-extinct states. In the revised version of the algorithm, this new state is drawn uniform randomly for *numRandJumps* times; after this it drawn from the current QSD representation. The current time, system state, rates and QSD representation are assumed to be accessible to the algorithm and are not explicitly passed in the methods.

```

1: initialiseSystem()           ► Initialise time, state, rates and QSD representation
2: while event < numEvents do   ► Keep simulating until numEvents reached

3:    $\tau \leftarrow \text{getNextTime}()$            ► Draw a time increment  $\tau$ 
4:   numCensusPoints  $\leftarrow \text{getCensusPointsPassed}(\tau)$ 
5:   if numCensusPoints > 0 then
6:     updateQSD(numCensusPoints) ► Update QSD representation with current state
7:   end if
8:   time  $\leftarrow \text{time} + \tau$            ► Update time

9:   state  $\leftarrow \text{getNextStates}()$ 
10:  if state = 0 then           ► If moving into the extinction state
11:    state  $\leftarrow \text{selectNewState}()$        ► Select a new non-extinct state
12:  end if
13:  updateState(state)
14:  updateRates()

15: end while

```

5.2.3 Compressions of the QSD

The main aim of this chapter was to consider possible compressions that would allow for the simulation of the QSD for larger systems. We tested three possible types of compression in comparison to the exact QSD. This was conducted for systems with a small number of patches for which the exact QSD can be obtained, by examining trends in accuracy as the number of patches was increased. The approaches were tested for three different distributions of patch areas, representative of different levels of system heterogeneity.

As explained above, the full QSD has dimension 2^n , and thus its storage rapidly becomes intractable as n increases. Under the stochastic SRLM model, all system states are reachable and thus the probability of any particular state is strictly greater than zero in the QSD (so we cannot simply ignore states with zero probability). Furthermore, *a priori*, we have no particular reason to expect that there will be equiprobable states in the QSD, so we cannot use an obvious grouping to reduce storage requirements. In general, in developing compressions, we must be willing to accept information loss (i.e. compressions of the QSD are expected to be lossy). The computational question arising is thus one of how to capitalise on statistical regularities in the the model that allow us to compress the information contained in the QSD while retaining the features of relevance for biological applications. In the simulation approach that we use, a representation of the QSD is required both as the output of the algorithm, and during the simulation. The latter situation is the case because whenever the system goes extinct, the chain is restarted by drawing form the current representation of the QSD. This means that any compression of the QSD influences the eventual output directly, but also during the simulation process when chains are restarted.

As it was unclear whether any existing compression algorithms (such as those used for compressing images or audio files) would be effective for this purpose, we decided to begin our search with compressions that are implicit in representations of the QSD in the ecological literature, and variants of these. Specifically, the ecological literature contains examples in which the QSD is represented by patch occupancy probabilities. This candidate compression corresponds to the assumption that patch occupancy is statistically independent (knowing that one patch is occupied does not tell us anything about the probability occupancy of any of other patches in the system). Since the colonisation rate of patches in the SRLM model is a function of the occupancy of surrounding patches, knowing that a particular patch is occupied changes our expectations about the occupancy of other patches, especially in the close neighbourhood of this patch. Assumptions of patch occupancy independence are therefore always a simplification; however, they allow us to store a compressed version of the QSD. There are different ways in which patch occupancy can be considered to be independent, and those implemented are described below.

The main aim of this chapter can be considered as one of identifying particular patch inde-

pendence assumptions that allow for sizeable compressions of the QSD representation while retaining an acceptable level of accuracy across systems with high and low levels of heterogeneity. In the first of the compressions, no patch occupancy dependencies are stored; in the second, the dependencies that are stored are pre-specified (dependencies are ignored once the number of occupied patches is specified); in the final approach, certain dependencies are captured, but in a manner where we do not have to pre-specify which these are, as they are determined by the data. The different compression algorithms are described below, and the corresponding storage format is shown in Figure 5.2 in an illustration with 5 patches; in the explanation that follows, we refer to rows and columns with reference to Figure 5.2.

1. *Full simulation*: As described in Section 5.2.2. The full QSD consists of a histogram of census point counts per system state and a total number of census points. These are stored in the form of an array of size $2^n - 1$ (or a hashmap up to this size), plus an integer value. This representation thus has a memory footprint of size 2^n (i.e. the space complexity is $O(2^n)$). States can be generated from the representation directly by drawing these with probabilities proportional to the state histogram counts. The number of operations required to write to the QSD representation is $O(n)$ and drawing a state from the representation is $O(2^n)$.
2. *Independent patches*: In this compression, at each census point, we increment by one an occupancy counter for each occupied patch. This algorithm treats occupancy of patches as fully independent, and has a memory footprint of size $n + 1$ (an array containing the census counter per patch, plus a count of total census points; the latter saves summing the occupancy counts every time a state is to be drawn from the representation). In other words, the space complexity is $O(n)$. The number of operations required to write to the QSD representation is $O(n)$ and drawing a state from the representation is $O(n)$. When a state is drawn from the compression, each patch is considered in turn and is chosen to be occupied with probability proportional to the occupancy count of that patch in the representation; rejection sampling is employed to exclude draws of the extinction state.⁵
3. *Independent patches by occupancy*: In this compression, we assume independence of patch occupancy once we have conditioned on the number of occupied patches. Storage consists of a vector of length n holding a count of the number of census points at which 1, 2, ..., n patches were occupied, and an $n \times n$ array of census counts per patch, conditional on the number of occupied patches. In Figure 5.2, columns correspond to counts in which the same number of patches was occupied, and individual entries

⁵This compression is closely related to the marginal distribution of patch occupancy probabilities; however, the representation itself is used to generate states when re-starting the chains after extinction so in general does not correspond to the marginal distribution of the full QSD.

to census counts per patch (the entries in the first column indicate that one patch was occupied at 3 census points; patch 1 once, and patch two twice). The total number of census points is also stored as an integer, and the memory footprint of the storage for this compression is thus $n + n^2 + 1$; the space complexity is $O(n^2)$. The number of operations required to write to the QSD representation is $O(n)$ and drawing a state from the representation is $O(n)$. Note that the final column in the array, representing full system occupancy, is not actually required as the counts are always equal to the corresponding vector value; however, it is included for algorithmic simplicity. States are drawn from the compression by first drawing a number of occupied patches from the row vector; rejection sampling is then used, drawing the occupancy state of each patch from the corresponding column, rejecting where the total number of occupied patches does not correspond to that drawn from the vector.

4. *Independent patches by cluster*: A clustering algorithm is used to cluster similar states (see below for details). Storage for each cluster k , up to a maximum of K_{max} clusters, consists of the number of census points for which the occupancy state vector is allocated to that cluster, and a counter per patch within the cluster. In the Java code, each cluster is an object consisting of an overall integer cluster count and a vector containing the patch occupancy counts within the cluster. This approach assumes that patch occupancy is independent within a cluster, and allows similar states to be grouped together where similar is defined over the individual patch occupancies (in contrast to the number of patches occupied in independent by occupancy). Including the storage of the total number of census points, this compression has a memory footprint of size $nK + K + 1$ where $K \leq K_{max}$ is the number of clusters used; the space complexity is $O(nK)$. Assuming that $K_{max} > n$, the number of operations required to write to the QSD representation is $O(K_{max})$ and drawing a state from the representation is $O(K_{max})$.

The number of operations in the simulation of the stochastic SRLM (without considering operations required to compute the QSD) scales as $O(n^2 + n)$ per transition in our implementation. For the simulation, we use the Gillespie Direct Method (Gillespie, 1977). As Ramaswamy et al. (2009) point out, the computational cost of simulation is dominated by the number of operations required to sample the next event and to update the rates once this has occurred. A more detailed explanation follows.

To sample the next event, we store all the rates out of the current state in an array. Because each transition affects a single patch, and each patch has only one possible transition, this array has length equal to the number of patches; using linear search, selecting from it therefore scales as $O(n)$. In our system, updating the rates dominates the computational complexity. Specifically, every extinction or colonisation event affects the colonisation rate of all unoccu-

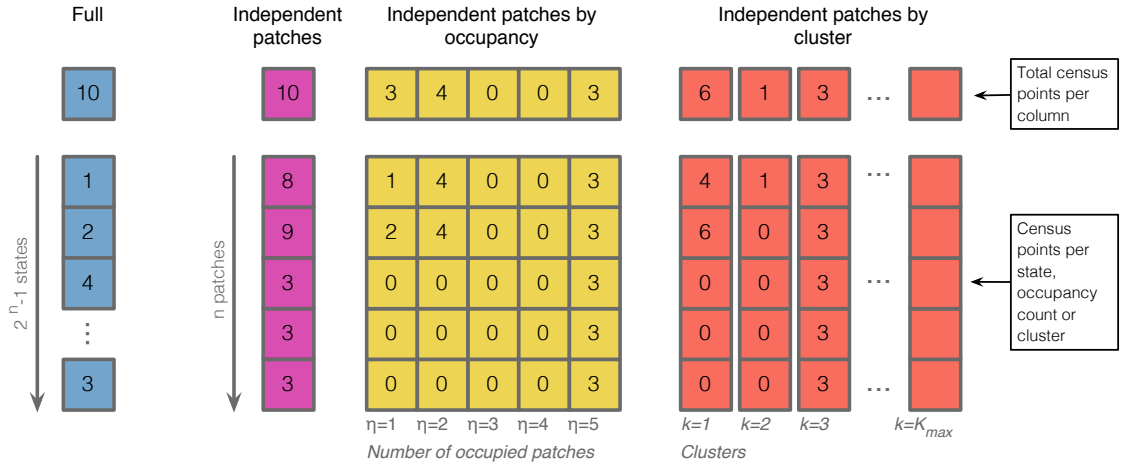


Figure 5.2: Illustration of system storage for a system with five patches after 10 census points for full storage and the three compressions, given states labelled such that system state one is [10000], for the hypothetical chain: [11000] [01000] [11000] [11000] [10000] [11000] [01000] (then extinction and uniform random jump to full state) [11111] [11111] [11111]. In independent patches by cluster, three clusters are used and the first records states [01000] and [11000]. Colours for the different forms of storage are used throughout the chapter.

pied patches (since the colonisation rate of any patch is given by the sum of the contribution from all other patches), so all colonisation rates need to be recomputed at each transition. The extinction rate of a patch is independent of the state of other patches and does not need to be recomputed on new events. (This simplification was not actually made in our code, however, because we were satisfied with the speed at which the code ran without making this improvement.) Denoting by $m \leq n$ the number of occupied patches, setting the rate for an unoccupied patch therefore requires summing over the contributions of each of the m occupied patches, giving a total of $(n - m)m$ operations of this type (each of which is, itself, $O(1)$); for occupied patches, we require m rate calculations. The worst case arises when the network is around half-full, i.e. for $m = (n + 1)/2$. The total number of rate computations required is then $(n^2 + n + 1)/4$. This means that the time complexity of the simulation scales as $O(n^2 + n)$.

The overall time complexity of the algorithm in 5.1 depends on the relative complexity of the drawing from the QSD representation (required only intermittently when the system goes extinct), the need to simulate enough state transitions to allow exploration of the state space, and the complexity of the simulation itself. Although the number of events required is not known *a priori*, we simulated for $2^n \times 1000$ events (the number of events thus varying according to the number of patches); if all states had been equally likely, this would have meant that on average, the simulator would have sampled each state 1000 times. For the sizes of systems used, this proved not to be prohibitive; for larger systems, a different stopping criterion would be required. This might be based on convergence of the QSD representation;

most applications are likely to focus primarily on the probability associated with common states, and convergence for these should happen much more quickly than for rare states. The overall space complexity is dominated by the storage requirements for the QSD, since the space complexity of the simulator itself scales as $O(n)$ (one needs to keep track of the current state of the n patches and the probability of a transition in each of them). In this study, we focus on compressions to reduce the space complexity of the problem.

Clustering algorithm

For independent patches by cluster, we use a clustering algorithm to construct a model of the probability distribution of system states by grouping together similar patch occupancy configurations. Our aim is to model the data in a manner that avoids us having to store the states in full.

The data produced by the simulator consist of a sequence of system states which can be written as patch occupancy vectors \mathbf{X}_t of length n with elements $X_{t,i} \in \{0, 1\}$ where i denotes the patch and t the census point counter. We begin by making the assumption that these vectors are drawn from a model in which similar states are grouped into clusters, denoted k . Each cluster is characterised by a cluster probability and a probability distribution over patch occupancy vectors. Specifically, each cluster k has a parameter π_k corresponding to the probability of selecting this cluster and a vector θ_k of patch occupancy probabilities associated with it. Within each cluster we assume independence between the occupancy status of individual patches: an occupancy vector is modelled as resulting from n independent Bernoulli trials, each with its own probability of success $\theta_{k,i}$. This assumption implies that although patches are independent within each cluster, they are not independent overall. In fitting the model, we wish to learn π and the patch occupancy probability vector θ_k associated with each cluster.

In the Bayesian clustering approach employed, we treat the unknown parameters π and θ_k as random variables. Specifically, we assume that the patch occupancy probability vector associated with the cluster θ_k is distributed as the product of beta distributions; for each patch, $\theta_{k,i} \sim \text{Beta}(a, b)$ (in which a and b are fixed hyper-parameters).⁶ Similarly, we treat the probability of each cluster as a random variable π assumed to be distributed according to a Dirichlet($\phi/K, \dots, \phi/K$) (in which the ϕ is the hyper-parameter). Note that sampling from $p(\theta|a, b)$ returns a vector of parameters for independent Bernoulli draws (one for each patch) while sampling from $p(\pi|\phi)$ returns a vector that constitutes a parameter of a multinomial distribution. In other words, we make the assumption that the data \mathbf{X} were generated by a

⁶The patch occupancy probability vector associated with the cluster θ_k is distributed according to the product of the individual patch occupancy probabilities because we assume patch occupancy independence within clusters.

process that proceeds as follows

1. Sample π from $p(\pi|\phi)$ (i.e. sample a multinomial from a Dirichlet)
2. For each cluster k , sample θ_k from the product of beta distributions $p(\theta_k|a, b)$
3. For each data point to be generated \mathbf{X}_t , where t is in $1..M$
 - (a) Sample its source cluster (a single observation drawn from the multinomial with parameter π) and set an indicator variable $z_{tk} = 1$
 - (b) For each patch i , sample $\mathbf{X}_{t,i}$ (draw once from the Bernoulli distribution parameterised by $\theta_{k,i}$).

In order to use the data to fit the model, we need to introduce the indicator variable z_{mk} which is 1 if the m th object originates in the k th cluster, and zero otherwise (together these values form the variable \mathbf{z} with dimensions $m \times k$). Using Δ to denote the vector containing all the hyper-parameters, and applying the model inference version of Bayes rule (without the normalisation constant, i.e. $\text{posterior} \propto \text{likelihood} \times \text{prior}$), fitting the model now corresponds to the inference problem

$$p(\mathbf{z}, \theta, \pi | \mathbf{X}, \Delta) \propto p(\mathbf{X}|\mathbf{z}, \theta, \pi, \Delta) \cdot p(\mathbf{z}, \theta, \pi|\Delta) \quad (5.2)$$

$$\propto p(\mathbf{X}|\theta, \mathbf{z}) \cdot p(\mathbf{z}|\pi) p(\theta|\Delta) p(\pi|\Delta) \quad (5.3)$$

The proportionality in line 5.3 holds due to the independence assumptions of selecting the cluster and drawing a patch occupancy state from it inherent in the model.

We wish to sample from the posterior defined in Equation 5.3. Gibbs sampling is appropriate for models of this kind as it is possible to compute the conditional distributions analytically. We note that due to conjugacy, θ and π can be marginalised from the conditional distribution required to sample \mathbf{z} and it is thus unnecessary to sample them directly. We denote the set of data up to and including the m th data point by \mathbf{X} and the number of data points allocated to cluster k as c_k . Now, assuming that we have already allocated all of the data points in \mathbf{X} to a cluster and that these cluster assignments are stored in \mathbf{z} , the probability of assigning the $(m+1)$ th data point to cluster k is given by normalising the expression

$$p(z_{m+1,k} = 1 | \mathbf{X}_{m+1}, \mathbf{X}, \mathbf{z}, \Delta) \propto \frac{c_k + \frac{\phi}{K}}{\phi + m} \prod_{i=1}^n \left[\frac{a_{ki}^*}{a_{ki}^* + b_{ki}^*} \right]^{X_{m+1,i}} \left[\frac{b_{ki}^*}{a_{ki}^* + b_{ki}^*} \right]^{1-X_{m+1,i}} \quad (5.4)$$

by dividing by the sum of over all clusters k . Note that $a_{ki}^* = a + \sum_t z_{tk} X_{t,i}$ and $b_{ki}^* = b + \sum_t z_{tk} (1 - X_{t,i})$.

A Gibbs sampler would then proceed as follows. We would initialise \mathbf{z} randomly, sampling data points and reallocating them to clusters according to the probabilities in 5.4. However, the size of the state space means that it is not possible to store the data points in each cluster explicitly in order to perform this resampling (and indeed, doing so would be contrary to our aims). As a result, we are obliged to use an algorithm where states are allocated to a cluster as they are generated by the metapopulation simulation model and store only the information required for the inference. We allocate patch occupancy state vectors to clusters by proceeding as follows.

1. The first vector \mathbf{X}_1 is allocated to the first cluster.
2. The remaining vectors $\mathbf{X}_2, \dots, \mathbf{X}_M$ are allocated to one of the K_{max} clusters following 5.4.

Note that in the case of empty clusters, $c_k = 0$ so the expression becomes

$$p(z_{m+1,k} = 1 \mid \mathbf{X}_{m+1}, \mathbf{X}, \mathbf{z}, \Delta) \propto \frac{\frac{\phi}{K}}{\phi + m} \prod_{i=1}^n \left[\frac{a}{a+b} \right]^{X_{m+1,i}} \left[\frac{b}{a+b} \right]^{1-X_{m+1,i}}. \quad (5.5)$$

During the clustering process, we need to draw from the model in order to generate a new state every time the metapopulation simulation goes extinct. To do this, we take the most recent version of the model, draw a cluster with probability proportional to c_k (ignoring clusters that are unused) and then draw a state on the basis as a set of independent draws of a Bernoulli distribution on the patches, with parameter $\frac{a_{ki}^*}{a_{ki}^* + b_{ki}^*}$. Rejection sampling over states is used until a non-zero state is drawn because the QSD only contains non-extinct states. The same approach is applied when we draw from the clustered model to generate an approximate QSD for comparison with the exact QSD.

5.2.4 Experimental setup

In this section we explain the experimental setup and parameter values. We begin by explaining how we set up the landscapes; however, the same set of landscapes is used as in Chapter 4 and the reader can thus skip the next two paragraphs. We then explain the parameter values tested.

The same set of landscapes was used for all the experiments reported in this chapter. For each number of patches considered, we first generated 100 sets of patch centres located within a 5×5 unit area according to complete spatial randomness (Hanski and Ovaskainen, 2000). Then, for each set of patch centres, we generated patch areas according to three distributions: all equal area, drawn from an exponential distribution and drawn from a log

normal distribution. Patch areas for all landscapes were scaled so that the summed area of patches was 6.25 square units, representing 25% of the total. For each number of patches ($n = \{3, \dots, 10\}$), this resulted in a set of 300 landscapes.

The total area covered by the patches was selected at 25% in an attempt to avoid heavy overlap between patches; following Ovaskainen (2003), no additional correction for patch overlap was conducted, and patches were allowed to extend beyond the 5×5 boundary. Scaling patch area was necessary because the rates of colonisation and extinction are proportional to patch area, so increasing the total area with the number of patches would have increased patch colonisation rates and decreased patch extinction rates, resulting in more persistent systems. We chose to investigate three different patch area distributions in order to verify that our results were robust to this factor (log normally distributed patch areas were included following Hanski and Ovaskainen, 2000). Before scaling, the rate parameter of the exponentially distributed patch areas was 1; for log normally distributed patch areas, the log mean was 0 and the log of the standard deviation was 1.

Our main aim was to compare the accuracy of the different compressions as the number of patches increased. In conducting the experiments described, we simulated for the different patch size distributions and also varied persistence by changing colonisation-to-extinction ratio as a robustness check. Specifically, the parameters $\alpha = 1$ and $c = 1$ remained fixed throughout while e was varied between 0.1 and 1.0 in steps of 0.1. Unless otherwise stated, we simulated for $2^n \times 1000$ events (the number of events thus varying according to the number of patches), giving an expected number of 1000 events per system state if drawn according to a uniform random distribution. We recorded the state with a census interval of 0.1 time units throughout. At the start of each simulation, the initial state was chosen uniform randomly among states; uniform random initiation states were also chosen to restart the simulation for the first few extinction events (more details are provided below). For independent patches by cluster, we used a baseline value of $\phi = 0.5$, and $a = b = 0.001$ unless otherwise noted. The small values of a and b mean that the data is very dominant (which is clear from the equations for a^* and b^*) and equal values mean that we start from the *a priori* position that patches are equally likely to be occupied as vacant.

To compare the different approaches, we represented the QSD in all cases as a vector. For the full simulator, this simply required the normalisation of the census count data; for the other approaches, the QSD was constructed by drawing a number of states from the compression corresponding to the original number of events simulated and normalising the resulting histogram⁷. During post-processing, we compared the vector representations of the exact

⁷The number of events was typically fewer than the number of census points during the original simulation so the raw number of counts per state was lower for the histograms constructed from the compressions than in the full simulation. However, as discussed later, increasing the length of the simulation and the number of draws appeared not to result in a systematic improvement when tested for independence by occupancy.

QSD to those of the normalised full simulation and those drawn from the compressed representations. To test the correspondence between two vectors, we needed to select a distance measure. A number of measures have been suggested for comparing probability distributions (Cha, 2007); however, we could find no study in the literature that considered whether these measures were comparable between vectors of different lengths, an important consideration since we wished to understand the performance of the different algorithms for different numbers of patches. We therefore tested a range of distance metrics (see Appendix D.1) for bias in measuring the distance between uniform random vectors of different lengths. The only measure among those tested for which the mean distance was not sensitive to the length of the vector was the Pearson correlation coefficient r (or more properly, the Pearson distance $1 - r$), and this was used for the majority of the comparisons in the remainder of this study. Post-processing to obtain distance statistics between vectors was conducted in Java. Figures were made in the statistical programming language *R* (R Core Team, 2012), using *ggplot2* (Wickham, 2009).

5.3 Accuracy of the full simulation

In Experiment 1, we tested the correspondence between the full simulated QSD and the exact QSD derived from the eigenvector calculations. Our findings show that for the SRLM, the algorithm proposed by de Oliveira and Dickman (2005) does produce an accurate approximation of the full QSD in reasonable time (i.e. not solely for an order parameter of the system, as considered by these authors). However, refinements were required in the case of systems with low persistence in which the exploration of the state space is highly dependent on initial conditions.

Some of the Figures in this section serve as illustrations. For these, a sample landscape with 5 patches was used, and this is shown in Figure 5.3.

5.3.1 Algorithm refinement for least persistent systems

Informal testing prior to conducting the full experiment showed that the algorithm proposed in de Oliveira and Dickman (2005) appeared to be effective for simulating not just the marginal distribution of the number of occupied patches, but also the full QSD. However, it sometimes failed when the system had low persistence because the full state space was not explored sufficiently thoroughly. Figure 5.4 illustrates two runs (a and b) of the algorithm for the landscape shown in Figure 5.3 and parameters leading to low persistence, compared with the exact QSD. The problem in the run shown in (b) appears to be to not visiting certain states at all in the simulator.

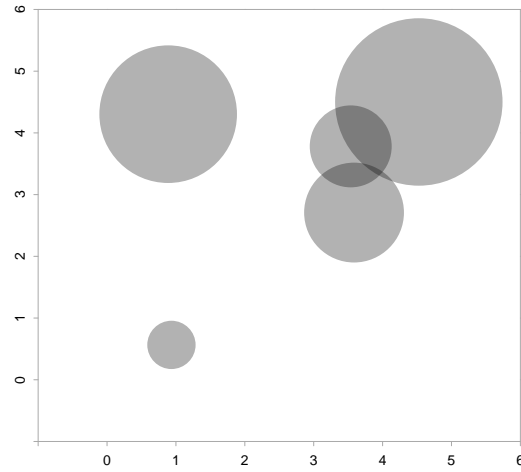


Figure 5.3: Sample landscape with five patches and log normally distributed patch areas.

In order to encourage better exploration of the state space, we increased the number of times that chains were restarted in a state drawn uniform randomly from all states. The number of random jumps required depended on the persistence of the system described by e and the number of patches n , and in all experiments described below, the number of times a uniform random state was chosen upon extinction was set to be $20 \times n \times e$.

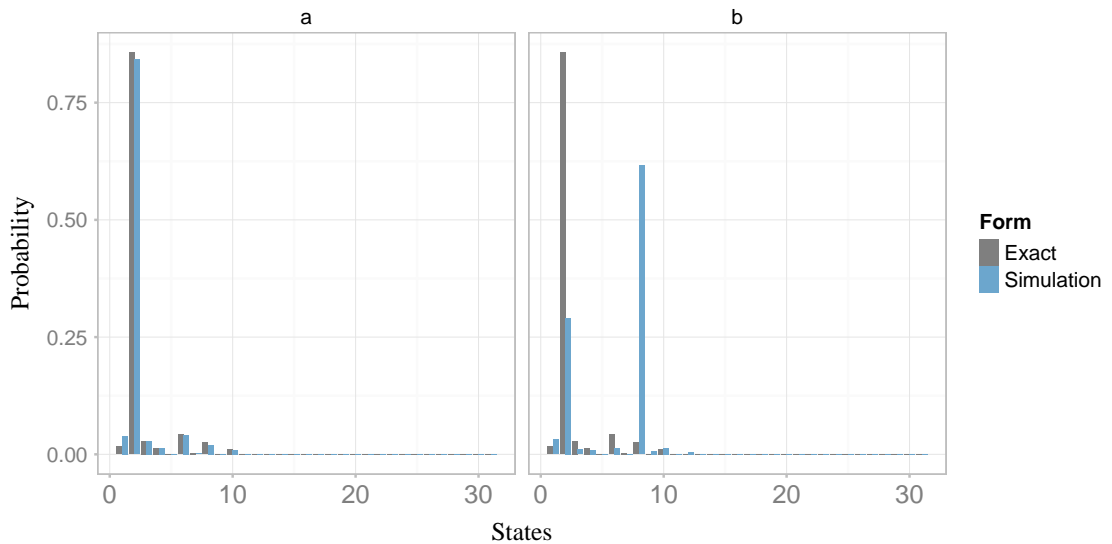


Figure 5.4: Two runs of the baseline algorithm for the landscape shown in Figure 5.3 and parameters giving low persistence.

5.3.2 Trend as n increases

We begin by providing an illustration of the accuracy of the algorithm (all discussion from here on uses the modified version of the algorithm described in Subsection 5.3.1 above). Figure 5.5 shows a histogram and scatterplot comparing the corresponding exact QSD and the full simulation method for a high persistence system for the landscape in Figure 5.3. For this example, the Pearson correlation coefficient between the two vectors is $r = 0.999898$, showing excellent correspondence. We choose a system with 5 patches for this illustration because it becomes difficult to interpret the histogram for the QSD with more patches. Although the particular parameter values illustrated give a better accuracy than the average for this number of patches (lower accuracy is achieved for systems with lower persistence across all patch size distributions), if we increase the number of patches to ten, this correlation is fairly typical, and thus it is representative of the accuracy for the larger systems which constitute our main interest in this work.

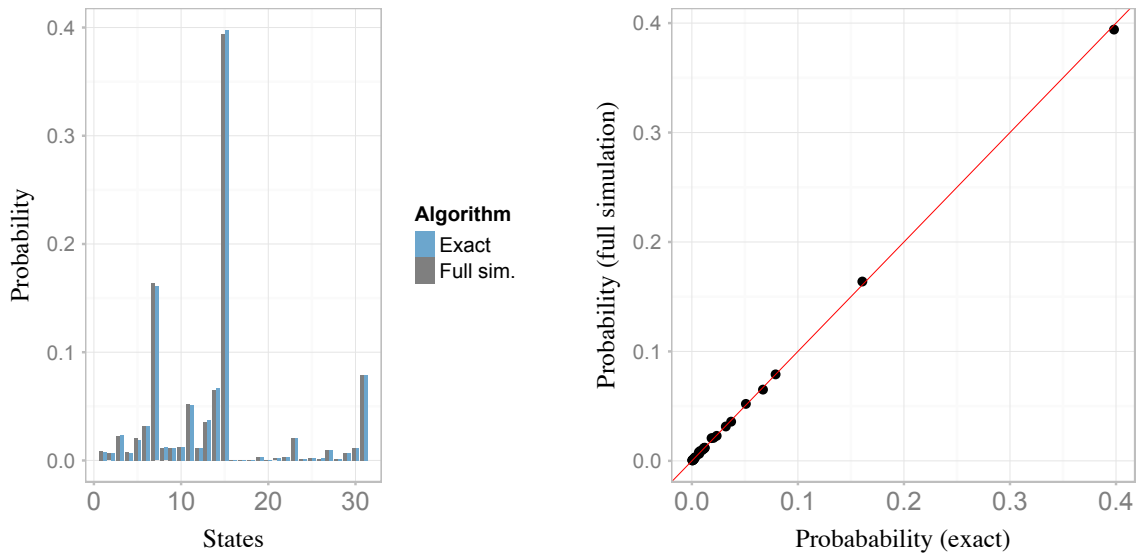


Figure 5.5: Left panel: State probabilities of exact QSD and full simulation (refined algorithm). Right panel: Scatter of state probabilities of full simulation against exact QSD. Both for the landscape shown in Figure 5.3 ($e = 0.1$, $\alpha = 0.5$).

Figure 5.6 shows the median Pearson correlation coefficient between the numerical QSD and the full simulation for a range of values of n and e . At least as far as $n = 10$, the full simulation improves in accuracy as n increases. The convergence of the full simulated QSD to the numerical QSD depends on the total number of events and the number of random jumps. Greater accuracy could therefore be achieved by simulating for longer, and particularly for low persistence, by including a higher number of random jumps. Among the 8000 simulation runs used to produce Figure 5.6, the lowest correlation coefficient was for a landscape with

four patches and log normally distributed patch areas and was $r = 0.8736$, leaving 23.68% of the variance unexplained; nonetheless the bottom 1% of correlations contained all values of n . For equal patch sizes, the correlations between the numerical QSD and the full simulation were slightly lower (overall median $r = 0.9991$ and lowest $r = 0.8332$, as opposed to overall median $r = 0.9997$ or 0.5991% of the variance unexplained for log normally distributed patch areas).

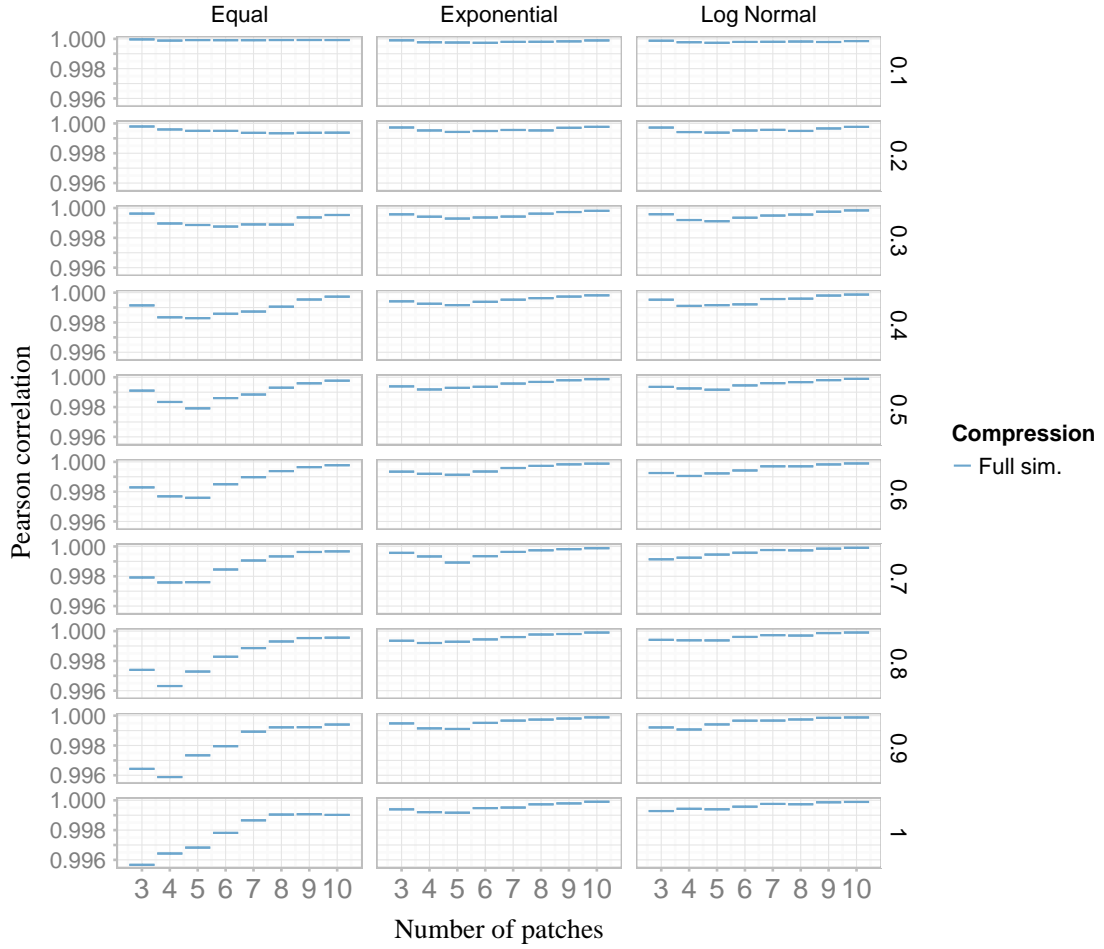


Figure 5.6: Algorithm accuracy in the form of median Pearson correlation between the true QSD and full simulation for different values of n and persistence e , shown on the right-hand axis (in all figures, persistence decreases as we move down the panels).

5.4 Accuracy of QSD compressions

The driving question in Experiment 2 is that of whether there is a compressed form of the QSD that provides an approximation to the QSD from the point of view of accuracy and that supports greater understanding of the system. As expected, all of the three compressions de-

scribed above (independent patches, independent patches by occupancy, independent patches by cluster) were less accurate than the full simulation. Although all compressions began to improve as n increased for some parameter values, there were also parameter sets for which all tended to become worse, and this trend was most consistent (although not most marked) for more diverse patch areas.

5.4.1 Relationship between compressions

We begin by comparing the accuracy (with respect to the exact QSD) of the simplest compressions - independent patches and independent by occupancy - to the accuracy of the full simulation. These compressions have a small memory footprint and are also straightforward to implement. Overall, Figure 5.7 shows that the full simulation performs much better than the two compressions, and independent patches is markedly worse than independent patches by occupancy. Aside from overall higher accuracy, independent patches by occupancy demonstrates similar patterns to independent patches, and we therefore only discuss the simpler compression in detail. For independent patches, the median correlation between the numerical QSD and the simulated QSD is $r = 0.9197$ over all parameter values. Nonetheless, for this compression algorithm, there were 15 landscapes with three patches and one with 4 patches for which the correlation coefficient between the numerical QSD and the simulated QSD was negative, with one correlation as low as $r = -0.7081$. All of these had relatively high e values ($e \geq 0.4$), showing that the largest errors occurred for small, poorly persisting systems. This may indicate that the number of random jumps at the beginning of the simulation was insufficient for these parameters in small systems.

Figure 5.7 also shows that as e changes, the importance of the assumptions of patch occupancy independence change for these compressions, and these assumptions are most problematic for intermediate values of e . When e is low, the landscape is relatively well occupied; colonisation of new patches is thus dependent on patches becoming available, a process that relies on patch extinction, which in turn is independent of surrounding patch status. As a result, the patch independence assumption is relatively unproblematic for low values of e (the median correlation for $e = 0.1$ across all n is $r = 0.9920$). When e is high, the system is in the ‘subcritical range’ and its dynamics are driven by patch extinction: in this situation, most events within the same chain (i.e. events not due to system extinction and the re-allocation of a new non-zero state) are independent patch extinctions, and the independent assumption is again not particularly problematic. In between these extreme values of e , patch colonisation forms an important part of the system dynamics and thus the patch independence assumption breaks down.

We now compare compressions with a comparable memory footprint: independent patches by occupancy and independent patches by cluster where the maximum number of clusters

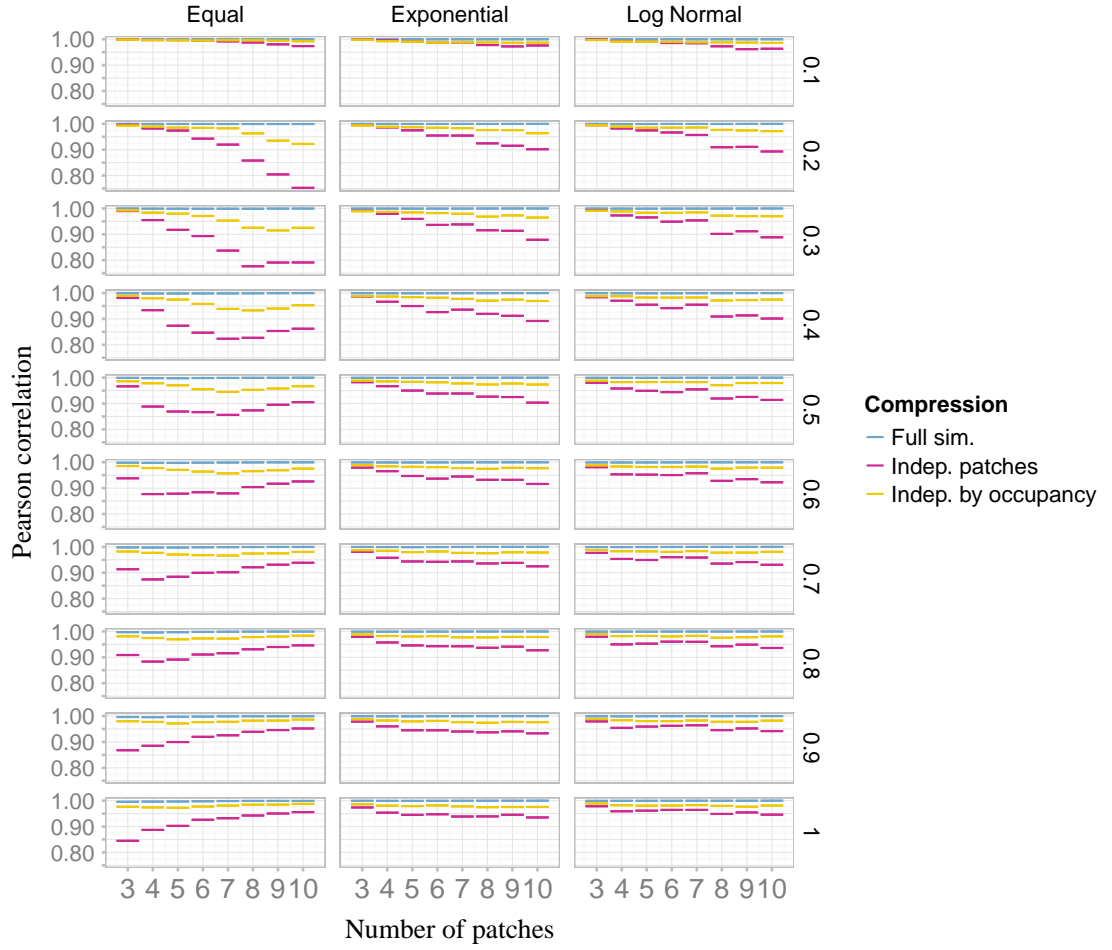


Figure 5.7: Algorithm accuracy of the full simulation, independent patches and independent patches by occupancy.

is equal to n . Figure 5.8 shows that for the same memory footprint, independent patches by occupancy is more accurate than independent clusters with $K_{max} = n$, and this advantage grows as n increases, especially for equal patch areas and log normally distributed patch areas for intermediate persistence.

In independent patches by cluster with $K_{max} = n$, all available clusters were always used. In order to verify that this effect was not simply due to the choice of the ϕ parameter that governs the tendency to start new clusters in the clustering algorithm, and which had been set to $\phi = 0.5$ originally, we tested the same approach with $\phi = 0.25$ (leading to a lesser tendency to use all clusters). We also wished to ensure that clusters were being used effectively. Because of the randomness associated with the generation of new clusters, it is possible for clusters to be created for uncommon patch configurations and attract only a small number of additional states to the cluster during the process of the simulation. Such clusters are, in effect, ‘wasted’ insofar as they store minimal information about the QSD. We therefore

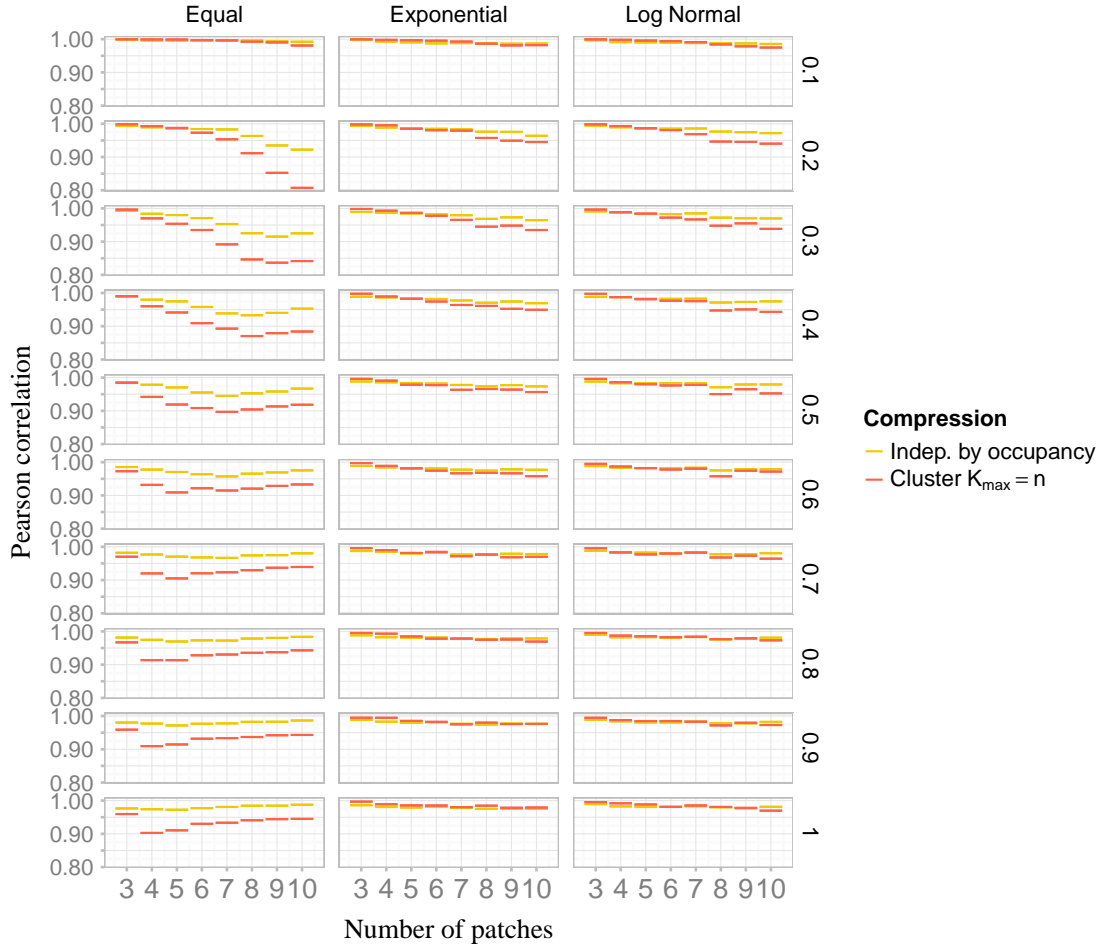


Figure 5.8: Comparison of compressions with similar memory footprint.

tested the compression allowing $1.25 \times n$ clusters, using the original ϕ value. The results from these tests are shown in Figure D.2 in Appendix D.2. These tests revealed that there was no systematic improvement due to these adjustments and independent clusters by occupancy still began to outperform clustering as n increased.

In order to further explore the accuracy of the clustering approach, we kept the original ϕ value, and increased the number of clusters to $K_{\max} = \{n^2, n^3\}$. Figure 5.9 shows that there were significant improvements by allowing up to n^2 clusters, and for some parameter values, small additional improvements when allowing n^3 clusters.

In order to understand the findings from this test, we plotted cluster usage (see Figure 5.10), finding that clusters were fully used for $K_{\max} = n$, that they tended towards full usage for $K_{\max} = n^2$, but that only a small proportion of the available clusters was used for $K_{\max} = n^3$, presumably explaining the small gains in accuracy with respect to n^2 clusters. As a result, we also tested additional ϕ parameters for n^3 to see whether we could achieve higher levels of accuracy by using more of the available clusters. We tested $\phi = 5$, but saw

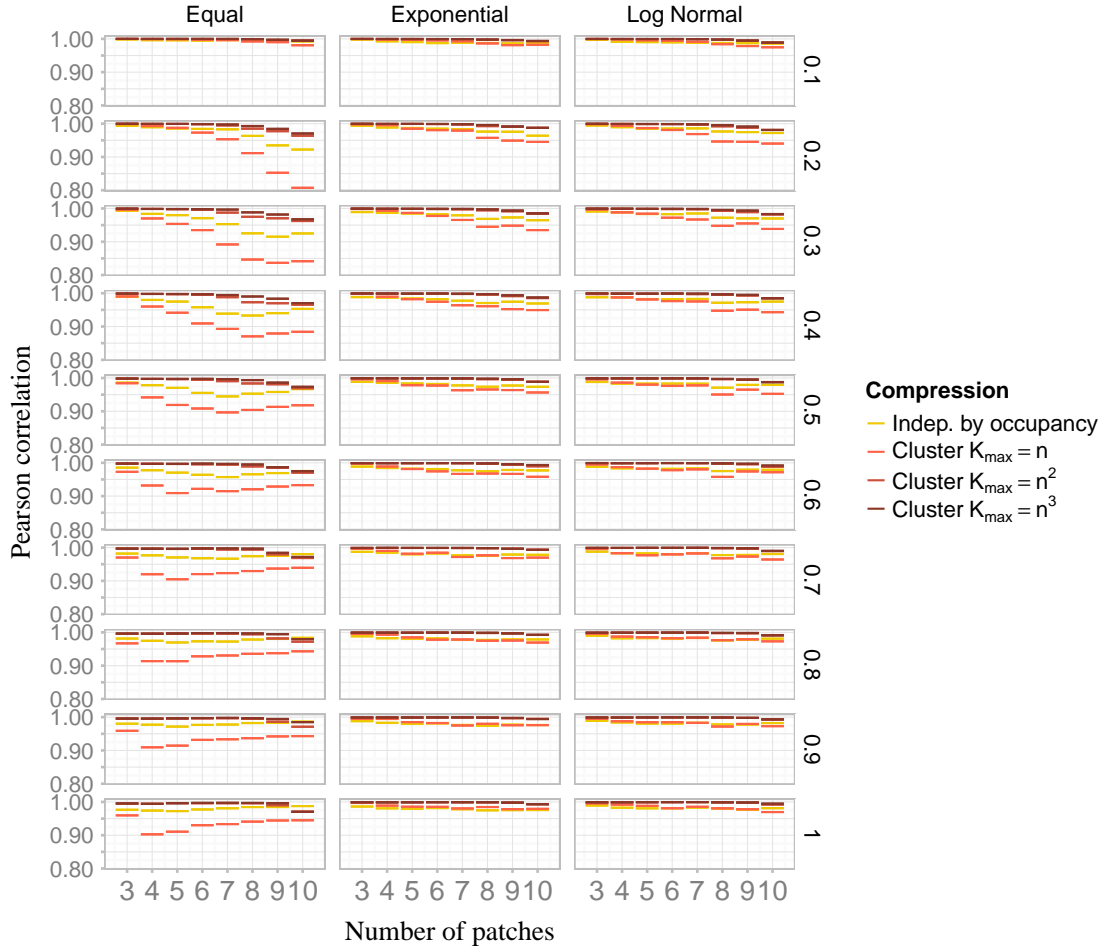


Figure 5.9: Comparison of adjustments to maximum number of clusters.

limited improvement, and as a result tested $\phi = 100$ and $\phi = 500$. Figure D.3 in Appendix D.2 shows the results of this comparison for exponentially distributed patch areas. Overall, correlations with the full simulator are very similar, and are still improving as the number of patches increases towards $n = 10$.

The memory footprint of the clustering approach with $K_{\max} = n^3$ grows as n^3 , in comparison with the full simulator, that grows as 2^n . However, for the largest number of patches tested $n = 10$, $n^3 = 1000$ is almost equal to $2^n = 1024$, so there is almost no saving from using the clustering. In addition, the computational overhead of calculating the clustering probabilities significantly reduces the value of this approach (the full simulator ran faster than the clustering approach for these parameter values), and selecting an appropriate value for ϕ is also difficult. In summary, for the model under consideration and the number of patches used in our experiment, the clustering approach appears to lose accuracy as n increases, unless K_{\max} is similar to the number of states.

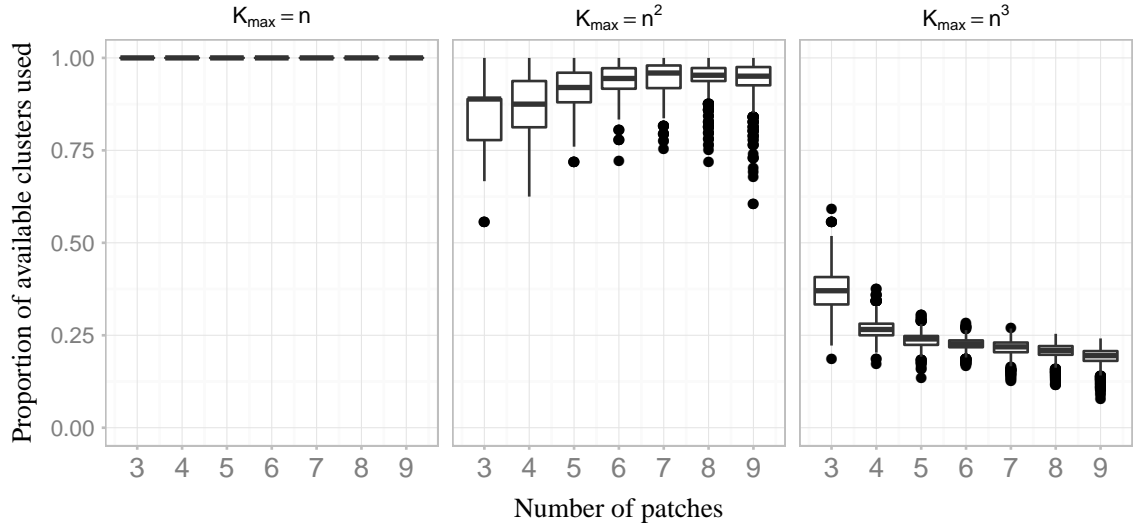


Figure 5.10: Proportion of available clusters used for landscapes with exponentially distributed patch areas.

5.5 Selection of a compression

In summary, none of the compressions tested is robust to increases in the number of patches and the experiments suggest that extrapolation beyond the number of patches tested is therefore difficult. All the compressions become less accurate as n increases towards $n = 10$, at least for some part of the parameter range, especially for intermediate persistence, and it is interesting to compare this situation to that of the full simulation which became more accurate for higher values of n . The compression that is least susceptible to this problem is independence by occupancy, where errors begin to reduce at higher n values in some parts of the parameter range, and may ultimately do so for all levels of persistence with sufficiently large numbers of patches. Although this means that it is not possible to provide guidance on a single best approach, one possible resolution to the problem is to select a compression according to what we hope to learn about the system. In this section, we provide guidance on preferred approaches according to this criterion and the number of states in the system.

5.5.1 Compression for overall accuracy

In some circumstances, we may be interested in obtaining the most accurate representation of the QSD. This might be the situation if we are hoping to draw from the QSD to generate starting conditions for a simulation, or if we wish to use the representation of the QSD for parameter estimation (e.g. if we have a real world dataset consisting of a set of states that are

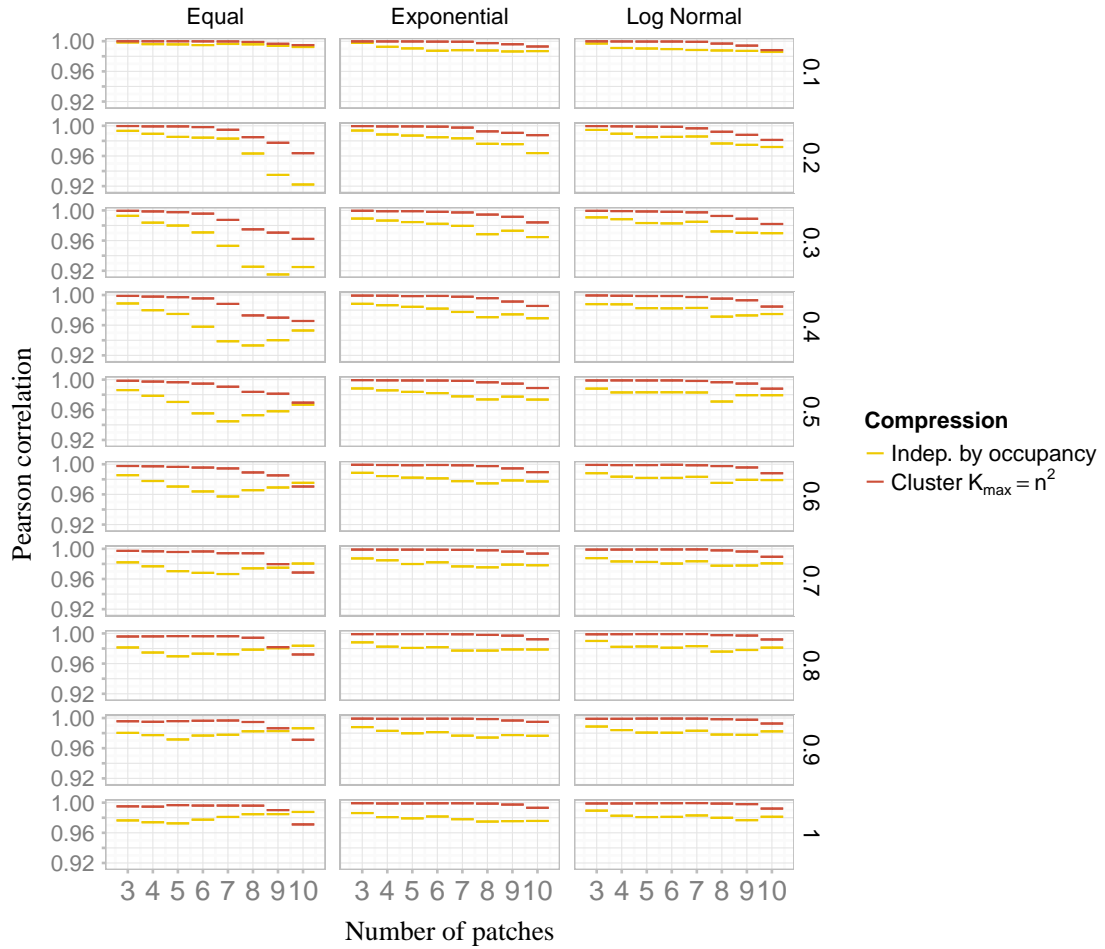


Figure 5.11: More detailed comparison of independence by occupancy with clustering with $K_{\max} = n^2$.

assumed to be drawn from the QSD and wish to conduct parameter estimation).

For small enough systems, the full QSD can be obtained numerically from the leading eigenvector of the generator matrix. This approach ultimately breaks down due to memory limitations. These first appear in relation to the storage required for the generator matrix, which has dimensions $2^n \times 2^n$. Sparse representations are possible since the generator matrix is sparse and becomes more so as n grows although whether these can be used to find eigenvectors depends on the programming language and implementation. This numerical approach may also break down due to numerical inaccuracy as the number of states becomes large and the eigenvectors more difficult to calculate without rounding errors. For example, Java supports arrays holding a maximum of 2^{31} elements (about 2.1 billion = 2,147,483,647). The full generator matrix for 15 patches has 1,073,741,824 entries (including 0th state) and is therefore the largest that can be stored in full in Java. (Although sparse representations may be possible, at least some packages, including the JAMA package, that can be used to calculate

eigenvectors do not have provision for sparse matrices.)

Once the numerical approach is no longer feasible, the full simulation can be used. Since the representation of the full QSD requires 2^n integer records (of census counts per state), it is possible to use this approach for larger numbers of states than the numerical approach, but this eventually breaks down because of memory limitations or because the maximum array size is reached. The theoretical maximum in Java is thus 31 patches, although in practice some elements may be allocated to array headers and size count, so the safe maximum is $n = 30$, provided computer memory allows. We note that it is preferable to store census counts rather than proportions of visits per state due to possible rounding error inaccuracies with the small values required by proportions.

Because the simulation time required for convergence of the algorithm may be related to the characteristics of the system (e.g. to the level of heterogeneity in patch area), the simulation should be run until longer runs result in limited change in the QSD properties. For example, one might test the correlation between the QSD when the simulation is run for E events and $2 \times E$ events, and between $2 \times E$ events and $3 \times E$ events and stop when the difference between the two correlations falls below a threshold value ϵ . In some systems, convergence may be slow and testing in this way is a useful check. Also, the system may appear to have converged to another distribution when the full state space is not explored. Therefore, testing different numbers of initial random jumps may also be important, especially for systems with low persistence.

Up until this point, we have been using the Pearson correlation as a measure of the accuracy of the QSD. This measure was selected because it is unbiased with respect to the number of states (see Appendix D.1), and because it captures the relationship between all state probabilities. It provides an overall measure of the relationship between two QSDs that allows us to derive the coefficient of determination (i.e. the proportion of variance explained); however, it is not always intuitive to interpret in terms of the magnitude of errors. In this case, other measures such as the root mean squared error (the mean state-wise error), maximum error or total error may provide more useful information. Figure 5.12 shows total error (square root of the sum of the squares of the error for each state probability) for the full simulator, independence by occupancy and the clustering approach with $K_{max} = n^2$. The total error decreases as n increases for both the full simulator and independence by occupancy. For example, for log normally distributed patch areas, the median total error for $n = 10$ is small (across all values of e it is 0.06072858). Obviously the root mean squared error per state is much lower.

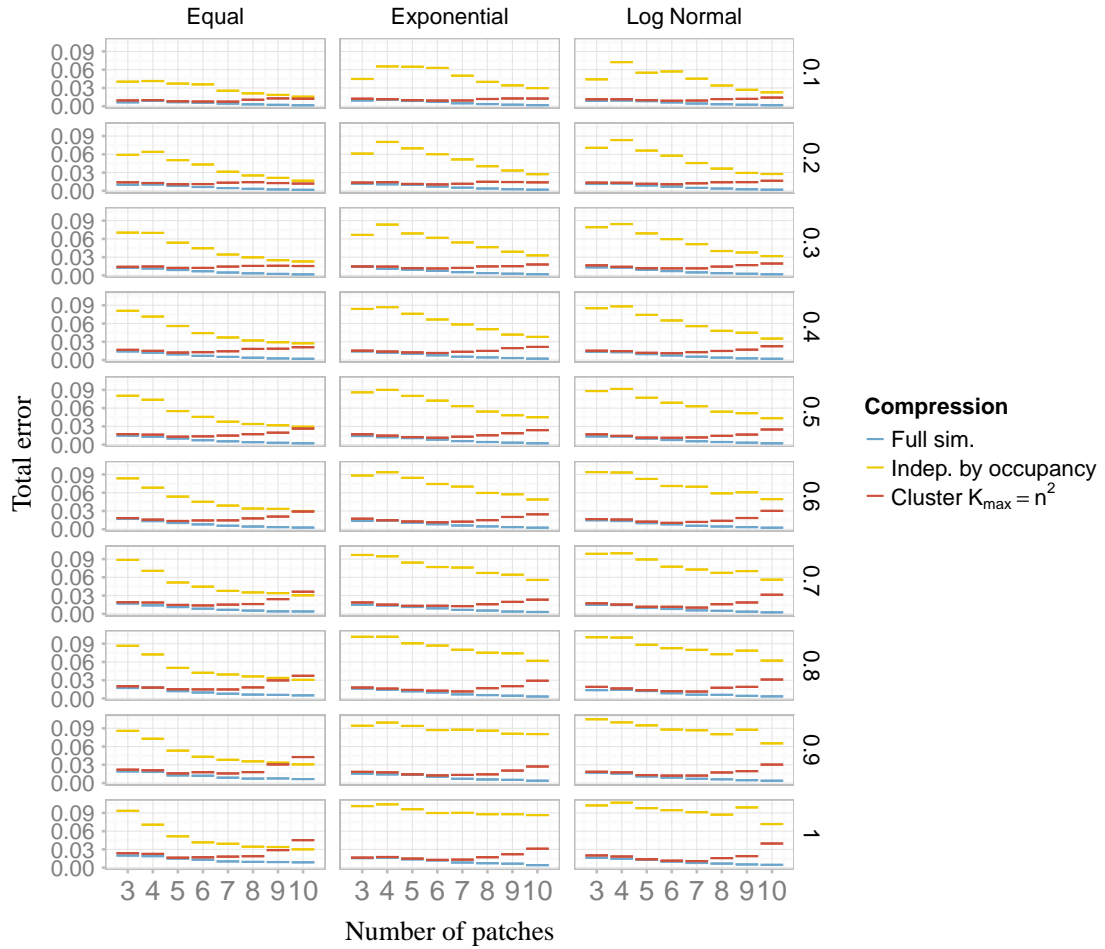


Figure 5.12: Median total error calculated as square root of sum of square errors.

5.5.2 Compression for extinction times and extinction paths

In some situations, we may be less interested in the overall accuracy of the QSD itself, and instead wish to use it to understand other system properties. One reason that the QSD is helpful is that the mean time to extinction from quasi-stationarity is well understood and can be found directly from the extinction rates from states in which only patch is occupied, and their state probabilities in the quasi-stationary distribution. It is therefore natural to ask whether any of the proposed compressions allows us to gain a good understanding of extinction times and how the approximations change as n increases.

Figure 5.13 shows proportional absolute error in mean time to extinction from quasi-stationarity for the full simulation and independent patches by occupancy. The independent patches by occupancy compression has the advantage of explicitly storing all of the near-extinction states. The percentage error for the full simulation reduces as the number of patches increases while the percentage error is more-or-less constant for independent patches by occupancy.

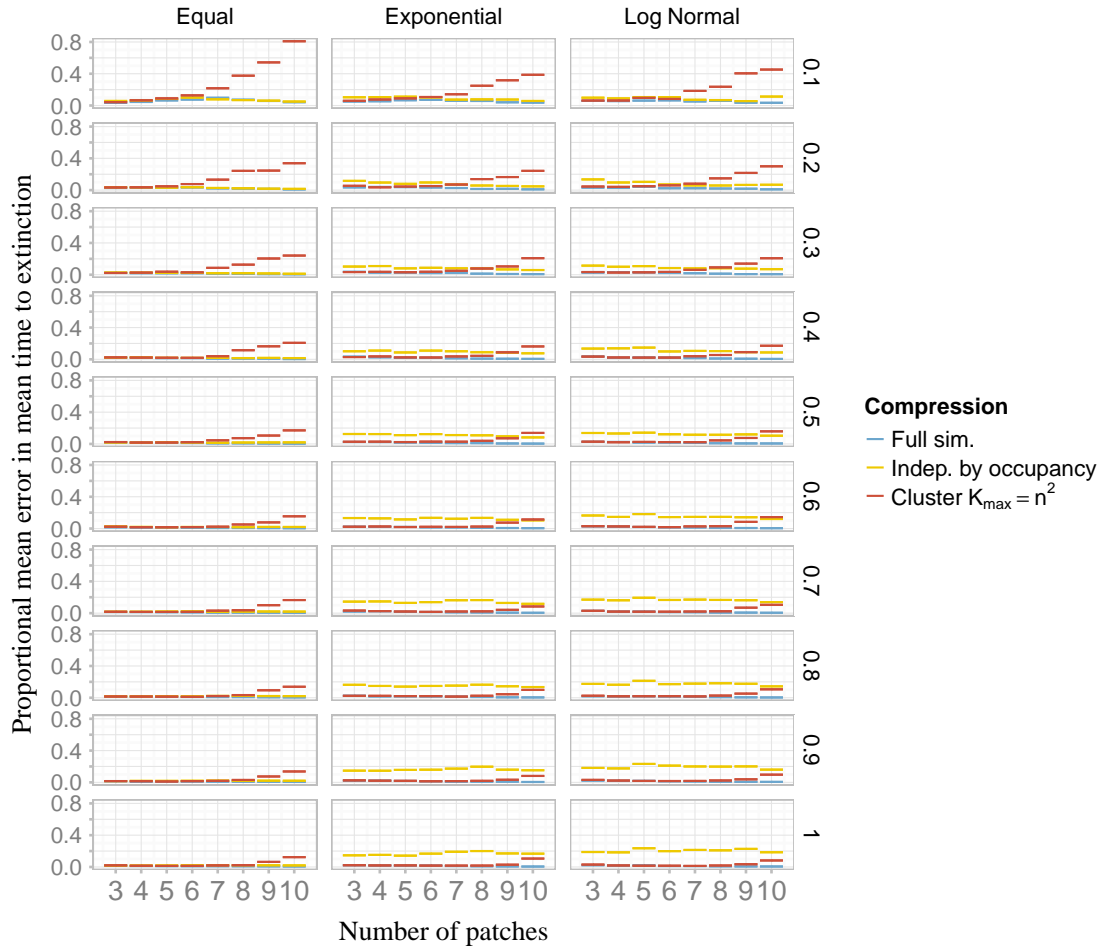


Figure 5.13: Proportional absolute error in mean time to extinction from quasi-stationarity.

As we reach higher numbers of patches, the error grows for the clustering algorithm with $K_{max} = n^2$, showing lack of robustness to increasing the number of patches and should be avoided for this purpose. Returning to independent patches by occupancy, although the median error for equal patch areas is only around 2%, for log normally distributed patch areas, it is around 14% (across all parameter values). Interestingly, the errors are smallest in the part of the parameter range for which the Pearson correlation coefficient demonstrated greatest divergence. For very high persistence ($e = 0.1$), the errors are relatively high, mostly because the near-extinction states will be visited only rarely in this situation, so longer simulation times may be required (i.e. a larger number of events should be simulated). If one wanted to investigate extinction times specifically, it may be best to continue simulating until each of the near-extinction states had been visited a pre-specified number of times.

To test this, we ran independence by occupancy for an order of magnitude more events and drew as many states as the original number of census points. This led to almost no improvement in the Pearson correlation coefficient (i.e. overall accuracy). However, it resulted in

improvements in the prediction of mean time to extinction from quasi-stationarity for the more persistent systems. Compared with the full simulation run for the original number of events, the longer simulation made better predictions for equal patch areas for $e = 0.1, 0.2$ and similar errors for other values of e . For exponentially and log normally distributed patch areas, the longer simulation performed slightly better for $e = 0.1$, with little improvement for lower persistence (and even possibly some worsening for the least persistent systems). For persistent systems and those with equal patch areas, independence by occupancy gives acceptable errors in estimation of mean time to extinction. For less persistent systems, the error does not grow as n increases, although overall errors are fairly large.

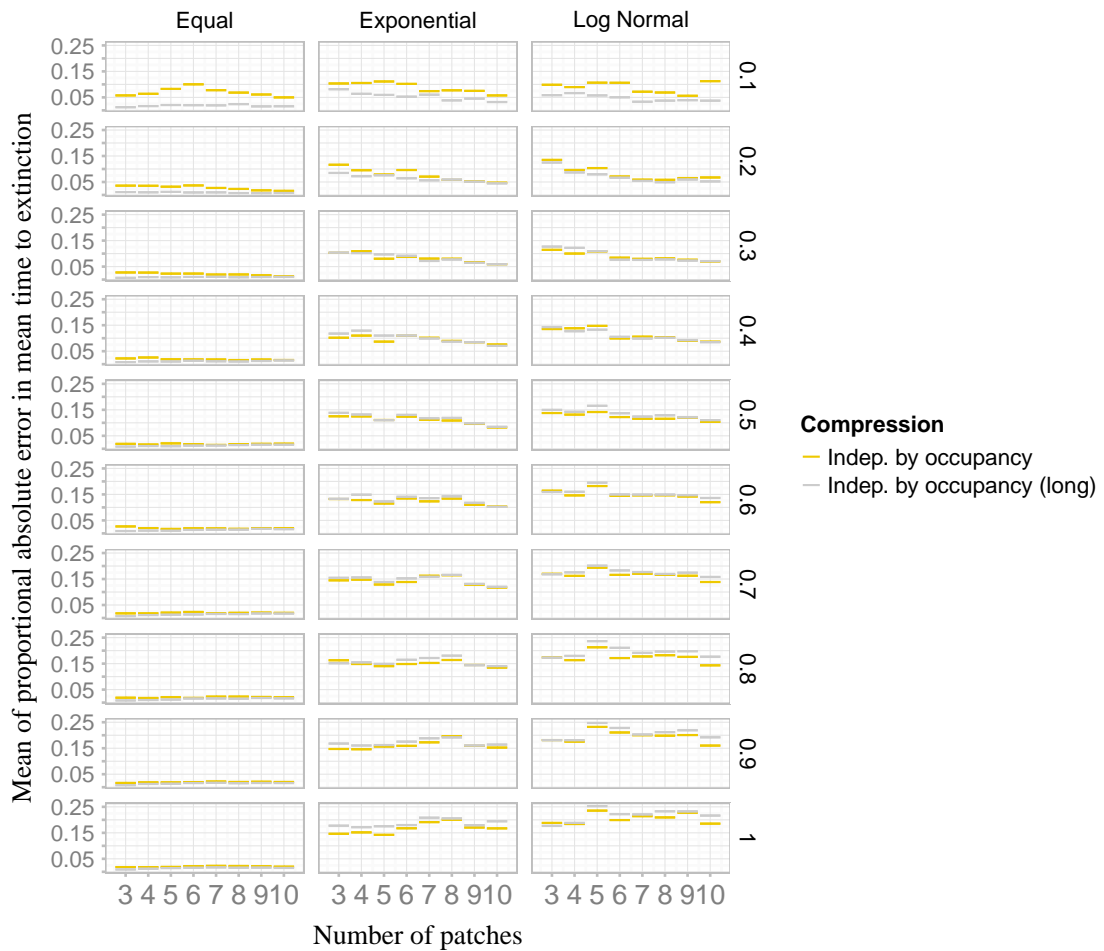


Figure 5.14: Mean of absolute error in mean time to extinction from quasi-stationarity, as a proportion of true T_m .

In addition to extinction times, we may be interested in likely extinction paths. For example, if we are interested in applying interventions to eliminate an infectious disease in a metapopulation, we might be interested in knowing where the infection is most likely to be when it comes close to elimination (e.g. is present in only one or two patches) as this might help us to ensure that resources are available in this location to finally eliminate it. For this,

independent patches by occupancy is also the preferred compression as it explicitly stores these states as separate. One can use the compression to investigate the most likely patch occupancy patterns close to extinction (with one or two patches occupied).

5.5.3 Compression for occupancy patterns

Another possible QSD application of interest might be in investigating patterns representing the most likely configurations in the endemic or long-term state for persistent systems. The full QSD is not very helpful for supporting this kind of understanding as it contains information on the probability of each state individually. The clustering algorithm, however, is ideal for this purpose as it allows us to investigate common configurations, and can be used at different levels of granularity (i.e. for different numbers of clusters). Figure 5.15 shows how this might be used in practice. In this example, the two most common clusters consisted of one shown in the left panel representing 47% of the census points in which most patches were occupied most of the time, and the small, isolated patch about 70%; and a second cluster (b) corresponding to 35% of the census points in which all patches apart from the small isolated patch were occupied. The next most common cluster only represented 16% of the census points, with the remaining clusters accounting for less than 5% of the census points (not shown).

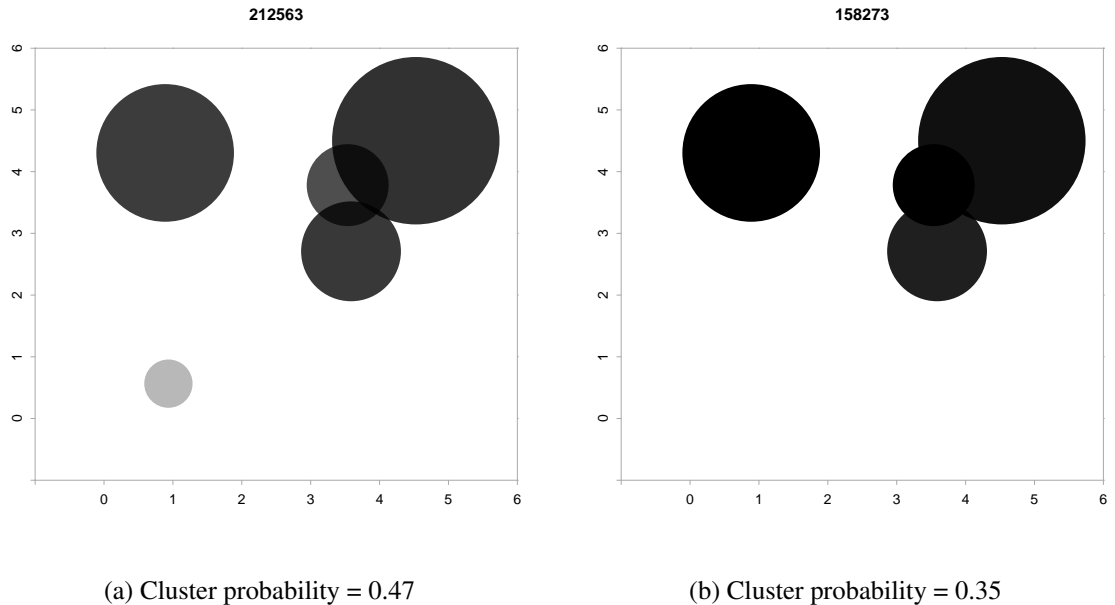


Figure 5.15: Probability occupancy (darker shading representing higher probabilities) of each patch in the two most common clusters in the cluster representation of the QSD for an example landscape with 5 patches and $e = 0.1$, and counts of the number of census points (out of 452537 census points during the simulation, using n clusters).

This approach is powerful for providing an overview of the system in a way that goes beyond simply plotting the occupancy probability of individual patches. A more nuanced picture can be gained by using a larger number of clusters; however, there is a trade-off between the level of detail that these represent and the insight that can be gained. Figure 5.16 shows that as more clusters are used, we begin to see that most of the time we expect the system to have either 3 or 4 patches occupied, but that all 5 patches is also common. The fourth most common cluster shows three patches occupied almost all the time and the isolated patch occupied only a small proportion of the time.

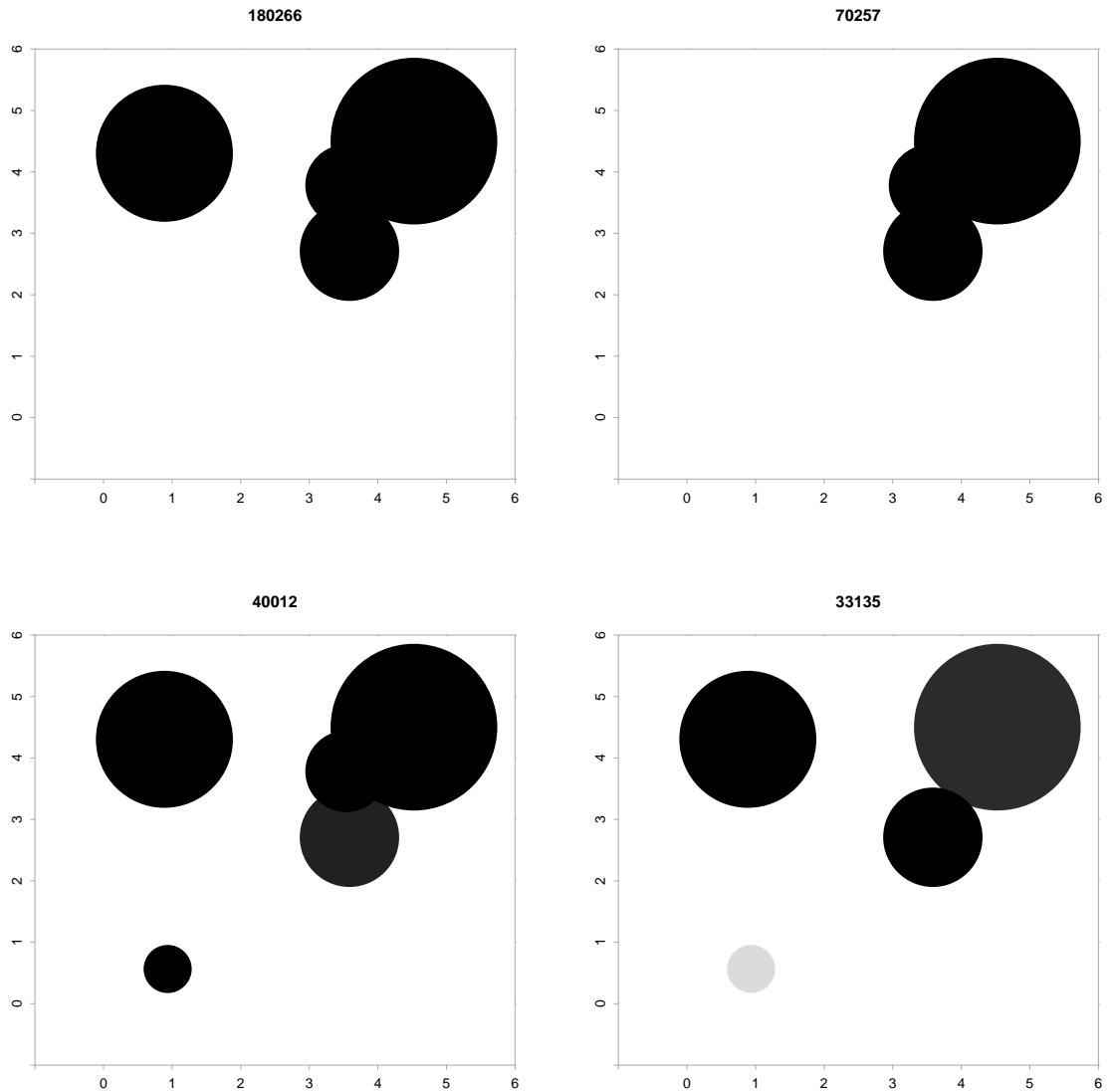


Figure 5.16: Probability occupancy of each patch (darker shading representing higher probabilities) in the two most common clusters in the cluster representation of the QSD for an example landscape with 5 patches and $e = 0.1$ with counts of the number of census points, using n^3 clusters.

5.6 Discussion

In summary, for small systems, the QSD can be found numerically from the eigenvector of the generator matrix. For slightly larger systems, it can be simulated using the full simulator. For larger systems still, our findings suggest that recommendations regarding an appropriate compression depend on the aims of the study.

For these larger systems, if one is interested in closely approximating the QSD, none of the compressions tested demonstrated good scaling with the number of patches for all levels of persistence when tested using the Pearson correlation; however, growth in the inaccuracy in independent patches by occupancy appeared to have slowed down, and had actually started to improve for lower levels of persistence. Furthermore, the median total error under this compression had started to decrease for all persistence levels by $n = 10$, suggesting that the Pearson correlation is much more sensitive to differences between two vectors than the median total error; in addition, the total median error values are small (e.g. 0.06 for $n = 10$, log normally distributed patch areas, averaged over values of e). Ultimately, whether this approximation is considered to be sufficiently accurate depends on the ecological questions one hopes to address.

If one is interested in mean time to extinction, the only algorithm for which errors were no longer growing as n increased to around 10 was independent patches by occupancy; however, the proportional error in the absolute value was relatively high. Nonetheless, there was systematic bias in the estimates of time to extinction in the direction of overestimation for this compression (see Figure 5.17); underestimates are rare and even the worst are small. We can therefore conclude that this compression provides reasonable estimates for mean extinction times for more persistent systems and that we can expect it to increasingly overestimate extinction times as the system becomes less persistent. In a disease ecology setting in which one attempts to control disease by making the system less persistent, this situation is unlikely to be particularly problematic: as one applies interventions to reduce the persistence of the system, predictions of mean time to extinction become progressively more conservative. As a result, although these predictions may not lead to an optimal allocation of intervention resources, and may lead to some resource wastage, their tendency to overestimate persistence should lead to more rather than less effective policy. In the context of conservation biology, these overestimates of persistence may be thought of as potentially more problematic; nonetheless, most systems for which such models are appropriate are likely to be relatively persistent compared with those considered here. This compression is also useful for investigating extinction paths since it explicitly encodes the near-extinction states. This may be useful in epidemiology to ensure that resources are available in the appropriate locations for elimination campaigns to be effective.

If one is interested in the endemic or long-run situation for more persistent systems, using

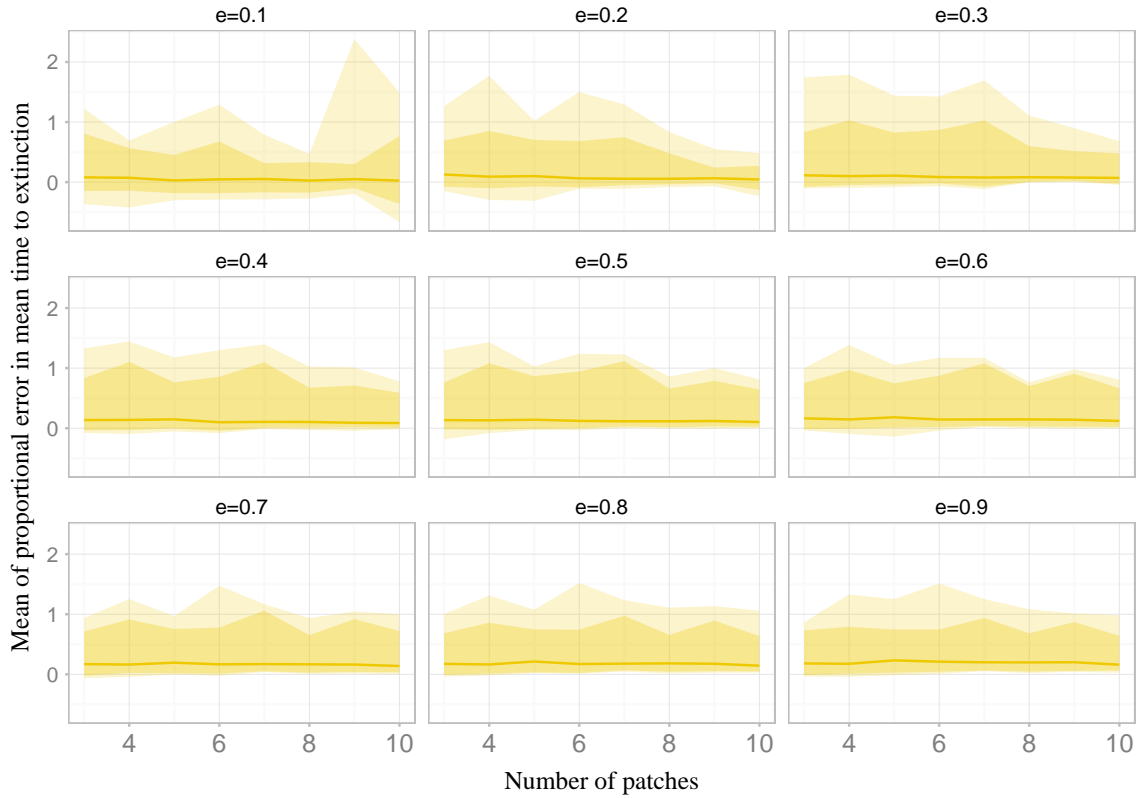


Figure 5.17: Median relative difference in estimates of T_m from independent patches by occupancy compared with those from the exact QSD (normalised by exact T_m , for log normally distributed patch areas) and showing 95% of the values (darker ribbon) and the full range (lighter ribbon).

a clustering approach may be more appropriate. The clustering algorithm used here allows the automatic grouping of common system states into clusters that are more amenable to visualisation than the full set of system states, a useful approach for gaining familiarity with the system. This may also be helpful for exploration of parameter space when attempting to fit models since one would expect most real world data on patch occupancy to fall within the most common clusters. For example, if particular patch occupancy combinations for a model are vanishingly rare under particular parameter values, observing these combinations repeatedly would suggest that a model fails to capture the real world; this information could easily be missed if we traced only the number of occupied patches. Information that can be used to compare model predictions of common QSD states and real world data will become available well before direct information on total system extinction. In fact, although datasets that consist of multiple time points will lead to better parameter estimates, the QSD should support model fitting based on a small number of snapshots for systems that are believed to be in the quasi-stationary regime. This obviously presents an important practical advantage over fitting methods that require time series data or turnover information

Specifically in relation to the deterministic SRLM, we note that the steady state solution of the differential equations for this model corresponds structurally to the first compression, independent patches. However, we have shown here that this compression provides a poor approximation to the full QSD. The question addressed in this chapter arose because we wished to initialise stochastic simulations in a state drawn from the stochastic equivalent of the deterministic steady state. Our results demonstrate that drawing such states from the deterministic steady state would be misleading: for example, extinction times from this compression are highly inaccurate (for the least persistent systems, the median error is around 25%, and errors for persistent systems $e = 0.1$), increase with n up to about an order of magnitude greater at 250% for $n = 10$, and something similar should apply to initial conditions drawn from the deterministic steady state. This conclusion is likely to generalise to other systems in which there is heterogeneity, and we thus caution against the use of the deterministic steady state for this purpose, at least for the SRLM and systems with important heterogeneities. This problem is likely to be even more acute if there is an Allee-like effect in the metapopulation since under these conditions, time to extinction from the deterministic mean may differ markedly from that from the QSD. Although the deterministic equilibrium is the same in the two cases, the QSD mass will focus much more heavily on the near-extinction states in the presence of an Allee-like metapopulation effect.

It would be valuable in future research to further investigate possible schemes for the compression of the QSD. For example, we might wish to combine the independence by occupancy scheme with that of the clustering algorithm, perhaps using clustering to characterise different occupancy states conditioned on the same number of occupied patches. It is currently unclear why independence by cluster was outperformed by independence by occupancy, and the extent to which this was due to the sampling process (according to which it was not possible to resample according to the usual Gibbs sampling approach) or rather due to characteristics of the particular system. Understanding the answer to this question may lead to better approaches. It would also be useful to consider the refinement of the original algorithm described here which includes re-starting the chain in a state drawn uniform randomly a number of times to see whether a less *ad hoc* approach can be developed, by exploring the convergence of different algorithms. For example, we begin with a QSD representation with zero counts for all states and initialising with some counts per state might be another way of encouraging state space exploration. Finally, it would be useful to have a more sophisticated measure of the correspondence between two QSD representations for the kind of system considered here. We have considered any two states as equally different from one another; however, for this kind of system, it would make more sense to consider, for example, states only differing in the occupancy of one patch, as more similar than states with very different numbers of occupied patches or no overlap in the patch occupancy.

5.6.1 Concluding remarks

The QSD can be used to investigate theoretical questions such as explanations of the core-satellite species distributions (Gyllenberg and Silvestrov, 1994) or the existence of mortality plateaux (Steinsaltz and Evans, 2004). It can be used in the context of model fitting, and is especially valuable where the system can be considered to be in the quasi-stationary regime and only a small number of system snapshots are available. It can be used to make predictions about common system states of a system that has reached endemicity, as well as paths to extinction and time to extinction from quasi-stationarity. These are valuable for practical applications such as disease control and conservation biology in which selection of the most appropriate patches for interventions is a key question of importance. To this end, the QSD of full states clearly provides more information than either the patch occupancy probability or the number of occupied patches.

In this chapter, we have provided a first attempt to simplify the simulation of the QSD in a way that can be employed with intermediate sized systems in which heterogeneity is an important consideration. The simulation approaches employed here are expected to be useful for ecologists, in part because they require minimal mathematical knowledge. In addition, if a simulation approach is already being used, it is easier to integrate into this a process for computing the QSD than to use a separate mathematical approach for this purpose. For example, one might choose to use a simulation approach to explore future scenarios, but need to test this with appropriate starting conditions; first simulating the QSD allows such initial conditions to be selected in a theoretically rigorous way.

Chapter 6

Persistence thresholds in related multi-state ecological and epidemiological models

Persistence thresholds have attracted attention as a possible way to establish the conditions required for metapopulation persistence, and have obvious practical value in both ecology and infectious disease contexts. Thresholds have typically been derived separately for different compartmental models. In this chapter, we first show how standard compartmental models for investigating persistence in epidemiology (namely the *SIS*, *SEIS* and *SIRS*) can be interpreted in the ecological and metapopulation context, and vice versa. In relation to the metapopulation structure, we focus on the original Levins and spatially realistic Levins models; for the former, we also apply both weak and strong Allee effects. We then explain the connection between the compartmental models by connecting them through the *SIIS* model that subsumes the specific models in the limit as we allow parameter values to vary. We use this connection to obtain persistence thresholds by first deriving these for the overarching model and then exploring how they change as parameter values tend to the limits that allow us to recover the specific models. The persistence threshold for several of these models has not been derived previously, while for the others, our approach for obtaining the threshold is new. We argue that investigating concepts of interest through the use of more general models, we can gain additional insight into the biology of different systems, and can quantify the effect of modelling decisions in relation to the choice between compartmental models.

6.1 Introduction

Predicting persistence under different conditions is important for the management of biological populations. Persistence thresholds have attracted attention as a way of formalising the distinctions between populations that are expected to persist and those that are not and thresholds have been derived for a range of classical and spatially-explicit models (e.g. Cobbold and Lutscher, 2014; Lloyd-Smith et al., 2005; Hanski and Ovaskainen, 2000). For conservation purposes, it would be helpful to be able to evaluate a critical amount of habitat required for long-term persistence of a particular species; in the context of infectious diseases, assessing the reduction in host population required for pathogen elimination is key to vaccination and similar public health campaigns.

In the previous chapters, we have used stochastic models to investigate persistence. In this chapter, one aim is to explore possible approaches to understanding persistence using deterministic models. We provide the derivation of novel persistence thresholds in the context of multi-state models. Nonetheless, perhaps a more important contribution of this work is in the more general framework formed by connecting model types and deriving mathematical expressions for important concepts using it. Viewed from this perspective, our threshold derivations illustrate this approach that relies on viewing compartmental models as a family in which some members are limiting models of a more general model. This approach helps us to connect and understand particular findings in relation to persistence thresholds, while raising new questions, and should be valuable in future work for guiding the selection of particular model structures.

Our discussion of persistence thresholds is motivated by a need to understand persistence in compartmental models in a range of application contexts. This kind of model is common in the epidemiological literature: in models of this type, individuals move sequentially through disease states. A simple example is the *SIS* model, according to which an individual moves from the state of being *susceptible*, to an *infectious* state, and then returns to the susceptible state. Compartmental models also appear in the ecological literature on metapopulation persistence (e.g. Hanski and Ovaskainen, 2000) where two states are typically employed, as well as in relation to system persistence in the context of successional dynamics with more states (Feng and DeWoody, 2004). Therefore we begin by relating models that have been employed in the epidemiological literature - in which these models have been well formalised - to the dynamics of metapopulations, viewed from both an epidemiological and an ecological perspective. That is, for each model, we consider different applications in epidemiological and ecological contexts.

The motivation for viewing persistence thresholds from the point of view not of separate models but families of models is based upon our view that this approach can lead to additional insights, and can be used to guide model selection. Although persistence thresholds

have been derived for models from these families, in which there may be more than two states, these thresholds have typically been addressed separately for the different models. Consideration of these different models in Roberts (2007) highlights the importance of making an informed choice of the disease classes to include in compartmental models and the relationship between these classes. Nonetheless, in his review of persistence thresholds, Roberts (2007) treats different disease models in different sections of the discussion.

In the following sections we begin by making the link between simple compartmental models of relevance for persistence in the infectious disease literature and those used in ecology. We briefly discuss the correspondence between the Levins model and Logistic model of population growth, extensions to a spatially-realistic context, and the role of Allee effects in these models. We then show the connections between these models, showing how the other models can be considered as limiting cases of the *SIIS* model. We then provide some background on existing work on persistence thresholds in these kinds of models, including work on metapopulation capacity. In the main section of the chapter, we provide examples of the derivation of persistence thresholds for *SIIS* models in a spatially-implicit and spatially-explicit context, with and without an Allee effect in the non-spatial version, and show how these can be used to derive the corresponding thresholds for related models. We conclude with a discussion of the implications of the approach and limitations in relation to the differences between deterministic and stochastic models. Some of the discussion in this chapter, especially in earlier stages, is likely to be familiar to those with a background in epidemiological modelling, although some of the links identified are likely to be new. Some aspects of the description of the Spatially Realistic Levins Model are also covered in earlier chapters, and the reader may progress through these rapidly.

6.2 Compartmental models

Compartmental models, in which individuals are modelled as being in one of a number of discrete disease states, are common in epidemiology. In this section, we introduce basic models and consider their application in an ecological context. In the literature on disease invasion and epidemics, individuals are typically modelled as progressing from a *susceptible* state (labelled *S*) to an *infectious* state (labelled *I*) in which they are able to infect other individuals; typically, they then progress to a *removed state* (labelled *R*) in which they can no longer be infected (either through death or recovery with immunity), and the full progression is denoted by the state labels: *SIR*. Sometimes, an additional *exposed* state (labelled *E*), precedes full infection, modelling a delay in the onset of infectiousness, resulting in an *SEIR* model. In contrast to epidemic models, models of persistence require the inclusion of

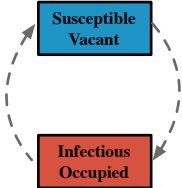
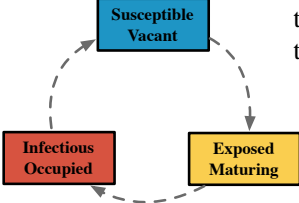
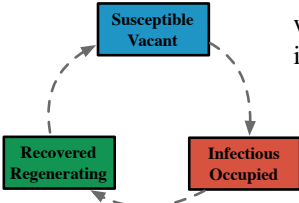
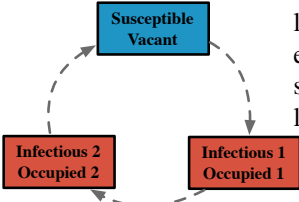
Model	Epidemiological individual example	Epidemiological metapopulation example	Ecological metapopulation example
SIS / Presence-absence 	Repeat infections, e.g. common cold <i>rhinovirus</i>	Diseases that infect only a small proportion of susceptible individuals, thus causing little or no population level immunity	Organisms that use habitat cyclically between presence and absence (and where habitat can be re-inhabited immediately following extinction)
SEIS / Maturation 	Diseases with incubation period before infectiousness occurs	Diseases with long incubation period: populations are infected long before being able to infect other populations. Population-level immunity not acquired (e.g. only a small proportion of susceptibles infected, causing limited immunity (e.g. rabies))	Social animals that require an establishment period before colonising new territories.
SIRS / Regeneration 	Diseases that procure waning immunity (e.g. influenza)	Diseases that infect a large proportion of a population, and procure immunity, but this is lost at population level (e.g. due to population turnover; childhood diseases such as measles)	Organisms that exhaust environmental resources; the environment must regenerate before recolonisation
SIIS / Successional 	Repeat infections (rather like <i>SIS</i>) but where each infection has two stages (e.g. a more and less infectious stage)	As <i>SIS</i> but with two stages of population infectiousness (e.g. a more and less infectious stage)	Successional dynamics in which a habitat is available for colonisation by different species in turn (e.g. grass-bush-tree)

Table 6.1: Illustrations of different compartmental models with, in the figures, epidemiological terminology (above) and ecological terminology (below).

longer-term processes, for example by allowing re-infection¹, and the compartmental models change from being linear to cyclical, modelling the processes by which individuals transition back to the susceptible state, giving rise to *SIS*, *SEIS*, *SIRS* and *SEIRS* models. In the simplest case, the *SIS* model, susceptible individuals alternate between susceptible and infected states, as might be the case for the common cold *rhinovirus*.

In making the connection between epidemiological and ecological models, it is useful to think about the epidemiological process from the point of view of the pathogen. In this view, individual humans constitute patches of ‘habitat’ for the pathogen: after being colonised, these are occupied until the host clears the disease as its immune system forces pathogen extinction. The dynamics can therefore be viewed as a colonisation-extinction process in which habitat patches (hosts) are either vacant (susceptible) or occupied (infectious). Conceptualisation as an *SIS* or a colonisation-extinction process is therefore a matter of perspective: the two models are mathematically equivalent.

The connection between *SIS* and colonisation-extinction models raises the question of the scale on which these models focus. Although compartmental models are typically applied to individual hosts in the epidemiological literature, we could model the infection status of populations in a similar manner: a population can be considered as a habitat patch and modelled as being susceptible or infectious (able to export infection to other patches). In other words, compartmental models can model habitat or host populations at different scales. For example, in an epidemiological context, the *SIRS* model describes a process according to which individuals or populations receive infection, become infectious and then recover to acquire immunity, finally returning to a susceptible state. In an ecological context, this corresponds to what we refer to as the ‘regeneration model’ describing a process in which a habitat patch is colonised, remains occupied for a time before local extinction occurs due to habitat deterioration, and the patch spends some time in a state of regeneration before returning to a vacant state ready for re-colonisation. Table 6.1 illustrates a range of models instantiated in an ecological context, where the focus is on the occupancy state of habitat patches, and in the epidemiological context in which the focus is on either individuals or populations. The table also includes the *SIIS* model that we shall consider in more detail in later sections as a model that subsumes the other models. However, the *SIIS* can also be interesting in its own right as a model that captures, for example, the situation when an infectious phase has two stages with different levels of infectiousness, possibly with different durations. This may arise if there is a acute infectious period and much longer chronic period because the host never entirely clears the infection (and thus never entered a recovered state), during which it can transmit at a reduced rate.

¹Longer-term processes may, instead, be modelled by incorporating demographic change in which new susceptible individuals emerge not through waning immunity but through reproduction (see e.g. what Hethcote, 2000, refers to as the ‘classic endemic model’).

6.3 A family of compartmental models

In this section, we provide background information on models used for the population structure for the compartmental models employed in later sections of the chapter. We begin with the logistic growth model, and move on to the metapopulation models used here: the Levins model, and the Spatially Realistic Levins Model (SRLM). The former is a standard model in the ecological literature and the latter is used in earlier chapters. The subsection on the SRLM is deliberately fairly detailed. Although the description of the model itself has been provided in earlier chapters, here we attempt to present the persistence threshold of this model in a much more intuitive fashion than in the literature to date. The reader may progress to Subsection 6.5.1, or indeed to Section 6.4, if this material is already familiar.

6.3.1 Logistic growth and Levins models

A basic model in the ecological literature is the logistic growth model. According to this model, growth rates are high when population sizes are small, modelling rapid growth due to the availability of resources, and fall to zero as resources are exhausted at higher densities. The rate of change of the number of organisms in the population N is described by the differential equation

$$\frac{dN}{dt} = \rho N \left(1 - \frac{N}{K}\right) \quad (6.1)$$

in which ρ denotes the intrinsic *per capita* growth rate and K the *carrying capacity* (the maximum number of individuals that can be supported by the habitat).

A model that is mathematically equivalent to the logistic growth model is the Levins model (Levins, 1969, 1970). This is a patch occupancy metapopulation model in which identical patches in an infinite network are modelled as either occupied or vacant. Vacant patches are colonised by occupied patches (or by individuals from occupied patches, but individuals are not explicitly modelled) at rate c , and go extinct or ‘recover’ at rate r (note that in this chapter, we use r - instead of e as used in other chapters - to denote the *per capita* extinction or recovery rate in line with our use of epidemiological terminology here). It is spatially implicit because connectivity is not modelled explicitly, nor is there any heterogeneity in connectivity between patches; as a result, there is no heterogeneity in patch colonisation or extinction rates. The model forms the basis of much of the work on metapopulation theory, for example providing a conceptual framework for understanding the relationship between habitat fragmentation and community persistence (Amarasekare, 1998). The model

is typically written

$$\frac{dp}{dt} = cp(1 - p) - rp \quad (6.2)$$

where p denotes the proportion of occupied patches.

Caughley (1994) and Amarasekare (1998) point out that this model is also a logistic model and that it is mathematically equivalent to the standard logistic growth model, where $K = 1 - r/c$ and $\rho = c - r$. As a result, we can write a re-parameterised version of the model in which

$$\frac{dp}{dt} = (c - r)p \left(1 - \frac{p}{1 - r/c} \right). \quad (6.3)$$

See Appendix E.1 for the full derivation.

6.3.2 Spatially realistic Levins model

In an adaptation of the original Levins model, Moilanen and Hanski (1995) developed the Spatially Realistic Levins Model (SRLM), used in Chapters 4 and 5 of this thesis. The SRLM differs from the original Levins model in that it captures heterogeneity in patch characteristics in both colonisation and extinction, and is no longer limited to an infinite number of patches, and it is in these senses that it is more realistic. Patch extinction and colonisation are governed by rates that depend on patch characteristics such as patch area or the connectivity of patches. In the general version of the SRLM, the dynamics of the probability occupancy of patch i is described by the differential equation

$$\frac{dp_i(t)}{dt} = C_i(\mathbf{p}(t))(1 - p_i(t)) - E_i(\mathbf{p}(t))p_i(t) \quad (6.4)$$

The SRLM incorporates additional heterogeneity in relation to original Levins model, although this comes at the expense of increasing complexity. Although it has not been used very extensively outside of its original area of application, it has the potential to be used more widely. The model was intended to capture metapopulation dynamics in a fragmented landscape, and was first used to model the presence or absence of the Glanville Fritillary Butterfly in grassy meadows. However, it can be thought of as a model of other heterogeneous systems at a range of scales. Taking foot-and-mouth disease as an example, a ‘habitat’ patch might consist of a region, a herd of cows, an individual cow or even areas of the cow’s body (its mouth and feet). At each of these scales, each patch could be susceptible (vacant) or infectious (occupied). In addition, in the SRLM, patches can be heterogeneous in the characteristics that influence both infection rates (e.g. connectivity to other patches) and

extinction rates (e.g. size). As a result, it would be possible to model differences in the probability of becoming infected of the different body parts. In another scenario, a patch might represent a group (e.g. an age-class) of individuals, where the value of the model is in its ability to capture the difference between stronger interactions within than between groups. The value of the SRLM is in the extent to which it captures heterogeneity in both the connectivity of patches and their individual characteristics, and the effects of these on colonisation and extinction processes². It is interesting to note that the notion of a metapopulation means that the question of whether a disease is epidemic or endemic becomes a matter of scale: an epidemic at a small scale might correspond to endemism at a larger scale.

Definition of metapopulation capacity

In this section, we explain background work on persistence in the SRLM. We attempt clarify the definitions in earlier work as well as making more explicit the connections between the definition of metapopulation capacity, the intuition behind its interpretation in relation to persistence, and its calculation. In the penultimate paragraph of 6.5.1 below, we also clarify the relationship between the metapopulation persistence capacity λ_M , the metapopulation invasion capacity and the next generation matrix interpretation of the basic reproduction number R_0 .

Persistence in the SRLM is related to metapopulation persistence capacity, denoted λ_M , defined as a measure of the capacity of a fragmented landscape, comprising patches, to support the long-term persistence of a species in the absence of external imports (Viana et al., 2014). It can be considered as a measure of effective habitat availability (ecology) or host abundance (epidemiology), weighted to account for factors such as patch or local host population sizes and connectivity, that influence extinction rates within patches or populations and transmission between them. Here we attempt to provide an intuitive explanation of the measure and its relationship with the persistence threshold. Given the general SRLM in Equation 6.4, and assuming that we can write the colonisation terms $C_i = cC'_i$ and $E_i = rE'_i$ for some $c, r > 0$, Ovaskainen and Hanski (2001) defined the measure as

$$\lambda_M = \sup_{\mathbf{p} \in \Omega} \left(\min_i \left[\frac{(1 - p_i) C_i(\mathbf{p})}{p_i E_i(\mathbf{p})} \delta \right] \right) \quad (6.5)$$

in which Ω represents the set of vectors \mathbf{p} and $\delta = r/c$. One can read the expression as follows: λ_M is the supremum over all possible occupancy vectors \mathbf{p} of the product of δ and the colonisation-to-extinction rates ratio in the patch i for which this value is minimised (i.e. for the patch that is growing slowest from this occupancy vector).

²Network models are often used in the epidemiological literature to capture heterogeneous contact processes; however, they less commonly include heterogeneities in recovery (disease extinction) processes. See e.g. Wilkinson and Sharkey (2013)

The parameters c and r are intended to represent species-specific characteristics and C'_i and E'_i are functions that describe colonisation and extinction as a function of the landscape characteristics. In order to see the value of extracting the parameter δ , we can write the expression as

$$\lambda_M = \sup_{\mathbf{p} \in \Omega} \left(\min_i \left[\frac{(1-p_i)}{p_i} \frac{c C'_i(\mathbf{p})}{r E'_i(\mathbf{p})} \frac{r}{c} \right] \right) \quad (6.6)$$

in which the c/r cancels with the r/c and we are left with an expression for the metapopulation capacity that is independent of the characteristics of the particular species, thus constitutes a measure of the landscape. Thus, unless the functions C'_i or E'_i are a function of other species-specific parameters, this formulation allows for the assessment of the landscape capacity independently of the individual species in question. Viana et al. (2014) follow Frank (2005) and define a similar measure

$$\lambda'_M = \sup_{\mathbf{p} \in \Omega} \left(\min_i \left[\frac{(1-p_i)}{p_i} \frac{C_i(\mathbf{p})}{E_i(\mathbf{p})} \right] \right). \quad (6.7)$$

Now, λ'_M is species-specific in that it constitutes a measure of the metapopulation capacity from the perspective of a particular species. However, it has the advantage of being easier to interpret.

In relation to the work in the current chapter, metapopulation capacity is of interest because it allows us to define a persistence threshold: if for a specific landscape, $\lambda_M > \delta$, then a population with species-specific parameter δ persists; otherwise it goes extinct. In the case of λ'_M , we have instead the persistence condition $\lambda'_M > 1$. The persistence criterion can now be written in the following way: a species persists in the metapopulation if and only if there is a probability occupancy vector $\mathbf{p} \in \Omega$ such that for all i , the colonisation-to-extinction ratio from \mathbf{p} is greater than one, i.e. $\exists \mathbf{p} \in \Omega$ s.t. $\forall i, \frac{(1-p_i)C_i(\mathbf{p})}{p_i E_i(\mathbf{p})} > 1$. In other words, there must be a system state from which every patch is more likely to be colonised than go extinct. If we consider the differential equation again, this means that $(1-p_i)C_i(\mathbf{p}) > p_i E_i(\mathbf{p})$, i.e. $\frac{dp_i}{dt} > 0$, that is, that occupancy probability is increasing for every patch. Put another way, there must be a \mathbf{p} such that the system is growing for all i . The measure λ_M therefore provides us with information on the size of the excess colonisation rate (in relation to extinction rate) of the weakest patch when we start from the conditions in which the probability occupancy of this patch is growing fastest. So λ_M is a measure of the fastest rate of growth of the probability occupancy of the weakest patch, across different initial conditions \mathbf{p} . In this sense, it provides a measure of the rate at which a particular system can tend towards a non-trivial solution.

We now explain the calculation of metapopulation capacity for simple SRLM models. For systems in which the rate at which a patch is colonised is given by a linear combination of

the colonisation rates from the donating patches, λ_M can be found as the leading eigenvalue of a matrix M . In order to write a matrix M , we need to be able to express the rate at which a patch becomes occupied as a linear combination of the contribution from other patches, because the matrix represents a linear operator. The matrix has elements $(m_{ij}) = \frac{C'_i(\mathbf{p})}{E'_i(\mathbf{p})} = \frac{r}{c} \frac{C_i(\mathbf{p})}{E_i(\mathbf{p})}$. This also suggests a relationship between λ_M and the eigenvalues of matrix M . Specifically, the eigenvalues of the matrix relate to the rates of change in different directions, the largest eigenvalue tells us the rate of change in the probability occupancy in the direction of fastest growth.

6.4 Links between different multi-state models

In the sections above, we described links between ecological and epidemiological models. Independently of whether they are used to model ecological or epidemiological processes, additional links can be made between the models with a different compartmental structure. These links are typically not made: epidemiological compartmental models are typically considered separately without making any explicit links between them (e.g. by considering *SIR* and *SEIR* as separate models). In the applied literature, the most appropriate model is usually selected for a particular disease and biological question of interest (Hethcote, 2000); in the theoretical literature, models are compared as discrete entities and either discussed separately or their properties compared (see e.g. recent discussions in Melesse and Gumel, 2010; Sun et al., 2012). Curiously, there seems to be little or no acknowledgement in the literature of the nested relationship of these models.

In order to explain the nested nature of the different compartmental models discussed here, we start with the *SIIS* model shown in Table 6.1. Using the terminology from epidemiology, in this model, there are two infectious compartments through which organisms progress sequentially (i.e. once infected, organisms must pass through the first and second of these in turn). This model allows us to model different levels of infectivity at the different stages of disease progression or might represent the effects of symptom evolution that cause different socialisation patterns that impact transmission at different disease stages (applications in metapopulation and ecological contexts are described in Table 6.1). Despite these possible applications, *SIIS* models are relatively rare in the epidemiological literature. Their consideration appears to have remained largely at the level of mathematical theory (see e.g. Melesse and Gumel, 2010, for a variant of the *SEIRS* model with an arbitrary number of infectious compartments), or is associated with studies of stochastic systems using the ‘method of stages’ approach (see e.g. Lloyd, 2001; Conlan et al., 2010), in which multiple infectious stages serve to transform the distribution of waiting times in the infectious compartment to give a more realistic distribution (Cox and Miller, 1965).

In the sections that follow, we show how *SIIS* models with two (or more) infectious compartments can be employed in an additional way, to provide insight into the relationship between simpler models, namely the *SIS* and *SEIS* and *SIRS* models. This is achieved by considering each of the *SIS*, *SEIS* and *SIRS* models as a limiting case of the *SIIS* model. In writing the *SIIS* model, we split the infectious state into two stages, I_1 and I_2 , with corresponding infectiousness c_{I_1} and c_{I_2} . Rates of transition out of the two infectious compartments are now labelled r_{I_1} and r_{I_2} , denoting recovery from each of the compartments. The model is illustrated in the left-hand panel of Figure 6.1. Because every individual who enters I_1 must also pass through I_2 , we need concern ourselves with only one of these states in our consideration of persistence. Focusing on the first infectious compartment, the dynamics of the proportion of patches in this state p_{I_1} for the *SIIS* Levins model can be written

$$\frac{dp_{I_1}}{dt} = (c_{I_1}p_{I_1} + c_{I_2}p_{I_2})(1 - p_S) - r_{I_1}p_{I_1}. \quad (6.8)$$

The *SIS* model can be recovered from the *SIIS* model by allowing the rate of transition between the first and second infectious state r_{I_1} (or between the second infectious state and the susceptible state r_{I_2}) to tend to infinity. The *SEIS* model can be recovered from the *SIIS* by allowing the rate of infection from the first infectious state c_{I_1} to tend to zero and the *SIRS* model by applying the same process to the rate of infection from the second infectious state c_{I_2} . The relationship between the models, using both the epidemiological and ecological terminology, is shown in Figure 6.1.

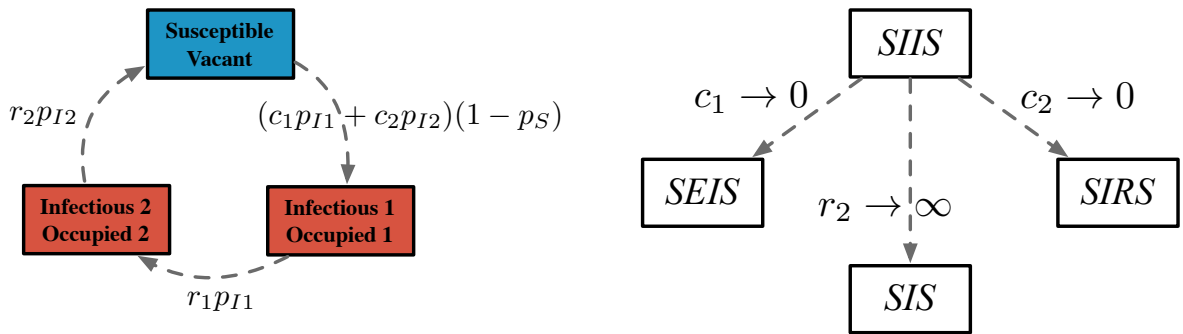


Figure 6.1: Relationship between compartmental models, in which c_{I_1}, c_{I_2} give the colonisation rates from I_1 and I_2 respectively; similarly for the recovery rates from these compartments r_{I_1}, r_{I_2} .

The idea of treating the *SIS*, *SEIS* and *SIRS* models as limiting cases of the *SIIS* forms the basis of the derivations in the following sections where it is applied to both the Levins model and the SRLM. We choose to examine the example of persistence thresholds within these models in line with the aims of the thesis. However, the main aim of chapter is broader than these specific models: we argue that by treating certain compartmental models as lim-

iting cases of more general models, we gain the ability to explore the connections between them in ways that might, firstly, help us to understand properties of the models (such as persistence), and that, secondly, can be used to guide modelling decisions about the appropriateness of particular compartmental structures for specific purposes.

6.5 Persistence thresholds and their problems

In this section, we outline two problems of persistence thresholds in deterministic models. The first of these relates to the application of persistence thresholds to stochastic systems, whereas the first relates to their usual mode of calculation. The first of these problems arises because of the binary nature of persistence thresholds: just above the persistence threshold, population sizes can be arbitrarily small. In stochastic systems, these arbitrarily small populations would be expected to go extinct very rapidly, and as a result, the persistence threshold does not provide a good measure of persistence.

In relation to the second problem, the typical approach to understanding population persistence in deterministic models consists of evaluating the stability of steady state solutions. It consists of finding the steady state solutions by solving for $dp/dt = 0$; the stability of these solutions is then found by considering the derivative with respect to p at the solution and if the derivative is less than zero, the solution is stable. This fact can then be used to establish the parameter range under which the solution is stable (see e.g. Case, 1999, for an accessible explanation). A related approach, based on the Jacobian of the next generation matrix (NGM), can be used in the case where there are a large number of states, or in metapopulation models including the SRLM (see e.g. Feng and DeWoody, 2004).

The application of this approach can be used at any steady state solution. In many models, including the two-state version of the Levins model and the SRLM, only two stable solutions are possible. In this case, the invasion and persistence thresholds are equal and the derivation of the ‘persistence’ threshold often actually proceeds by finding the invasion threshold by showing the conditions under which the trivial solution is unstable (instead of the conditions under which the non-trivial solution is stable), presumably because this simplifies the mathematics involved. Although this simplifies the mathematics, it does not necessarily lead to as much biological insight. For example, the *SIRS* model has only two possible stable solutions; as a result, we can evaluate the invasion threshold as this must be the same as the persistence threshold. For the *SIRS* model, with β the contact rate between susceptible and infectious individuals, and r the recovery rate, and ν denoting the rate at which recovered individuals return to the susceptible state, the set of differential equations governing the system

is

$$\frac{dS}{dt} = \nu R - \beta SI \quad \frac{dI}{dt} = \beta SI - rI \quad \frac{dR}{dt} = rI - \nu R. \quad (6.9)$$

Invasion occurs if the number of infectious individuals increases, i.e. if $\frac{dI}{dt} > 0$

$$\beta SI - rI > 0 \iff \frac{\beta S}{r} > 1. \quad (6.10)$$

At the beginning of an epidemic, one can assume that the number of susceptibles is approximately equal to zero, so we obtain the condition $\beta/r > 1$, equivalent to the expected threshold of $R_0 > 1$, and because there are only two stable solutions, we can also conclude that disease will persist for $R_0 > 1$. However, although the assumption of a completely susceptible population is sensible for disease invasion, it is clearly problematic in the case of long-term disease persistence, and deriving the persistence threshold in this manner provides little direct insight into the biological mechanisms of persistence. The result is particularly troubling because although as one would expect, the steady state prevalence is lower in the *SIRS* model than the *SIS* model, the persistence threshold is the same for the two models. Thus, although it is typically easier to find the persistence threshold by first obtaining the invasion threshold (e.g. because of the relative ease of substituting the trivial solution into the derivative), this functions only because of a mathematical property of the model, and little biological insight about persistence is gained. It would therefore be instructive to have an approach that provides more information about persistence and the counterintuitive findings such as the correspondence of the persistence threshold in the *SIS* and *SIRS* models.

The definition of persistence threshold provided by Hanski and Ovaskainen (2000) considers the persistence threshold directly in relation to the stability of the nontrivial solution, and is therefore more robust to situations in which persistence can occur even when invasion is not possible (e.g. when there is an Allee effect). Nonetheless, presumably for reasons of analytic tractability, most of the work carried out by these authors focuses on the SRLM without an Allee effect. In this situation, they show that the persistence threshold (corresponding to the existence of a non-trivial steady state) can be derived as the leading eigenvalue of a matrix, that they call a ‘landscape matrix’; in other situations, the threshold can be found numerically (Ovaskainen and Hanski, 2001).

6.5.1 Allee effects

First described in the 1930s, an Allee effect can be said to exist when there is a positive relationship between individual fitness and population size (or density) (Allee, 1931). This means that populations go extinct when rare, making initial invasion difficult and persistence

of small populations unlikely. As noted by Boukal and Berec (2002), the most commonly cited cause is that of the difficulty of finding an appropriate mate in low density populations of sexually-reproducing individuals. However, other causes are reviewed in Berec et al. (2007) and include inbreeding depression, tendency not to be pollinated, or the minimum group size required to raise offspring, search for food or avoid predator attacks (Boukal and Berec, 2002).

A distinction in the literature exists between strong and weak Allee effects (see e.g. Brassil, 2001). Strong Allee effects occur when there is a threshold population size or density below which population growth is negative. In the case of weak Allee effects, there is no threshold and growth rate is a positive function of population size, at least at small population sizes. From a mathematical perspective, a strong Allee effect means that the extinction state is a locally stable equilibrium point of the system for all parameter values (McVinish and Pollett, 2013a). Allee effects can also apply to birth or death processes, or both. For example, inability to find a mate affects birth rate while the existence of a minimum group size required to avoid predator attacks relates to death processes. Multiple effects may be present simultaneously. Berec et al. (2007) reviews possible interactions between Allee effects and their combined role in contributing to persistence.

Allee-like effects have also been discussed for metapopulations. In this context an Allee effect refers to a reduction in the colonisation capacity of patches when few patches are occupied. More precisely, the growth rate of the metapopulation due to exports from individual patches is an increasing function of overall occupancy at low levels of occupation (McVinish and Pollett, 2013a). Amarasekare (1998) reviews evidence supporting the existence of metapopulation level Allee effects. Zhou and Wang (2004) show that an Allee-like effect in a metapopulation can emerge from an imposed Allee effect at the local population level. While the inclusion of an Allee effect has a negative effect on persistence in general, Brassil (2001) provide an illustration of a model in which the effect is much stronger when applied to metapopulations than in the case of single, fully-mixed population. Further, McVinish and Pollett (2013a) show that habitat degradation can have much stronger effects on populations with a metapopulation level Allee effect than those without. In the case of metapopulations, Allee effects can be implemented as affecting colonisation or extinction processes, or both.

Although Allee effects have been discussed most commonly in a conservation biology context, they have also been invoked as being important in epidemiology. Most discussion relates to the effect of pathogens on host species subject to an Allee effect. In this context, the additional burden of disease can reduce numbers of a species sufficiently to cause extinction (Haydon et al., 2002; Hilker et al., 2009). Although such effects are rarely mentioned, Allee effects may also be present directly within pathogen and parasite populations. In sexually-reproducing parasites, many of the causes described in Berec et al. (2007) still apply. However, even in asexual pathogen populations, Allee effects may still apply. For

example, it is likely that they could be caused by reductions in genetic variation when population sizes are small. This effect could be compounded by the increased probability of gene fixation in small populations due to drift (this mechanism is suggested by its existence in other clonal organisms such as vegetatively reproducing plants Fischer et al., 2008). Other mechanisms in asexual pathogens could be those typically considered in the context of co-operation. For example, slime moulds can reproduce sexually in the amoebae state but also asexually through cellular budding and through sporulation mechanisms. In order to reproduce through spores, they need to cooperate to form a fruiting body, something that requires a minimum population size. Similar kinds of cooperative behaviours have been invoked in the case of *Pseudomonas aeruginosa*, a common disease-causing bacterium.

Boukal and Berec (2002) review different functional forms for incorporating Allee effects into population dynamic models. In later sections, we follow Amarasekare (1998) in the incorporation of a strong Allee effect, and Tabares and Ferreira (2011) for the incorporation of a weak Allee effect.

In addition to the persistence threshold, Ovaskainen and Hanski (2001) also discussed the invasion threshold. The metapopulation invasion capacity gives rise to the threshold condition $\lambda_I > \delta$ above which the trivial equilibrium state is unstable, meaning that a single small local population is able to invade an otherwise empty network. In cases corresponding to the Levins model, the invasion threshold is the same as the persistence threshold, and the approach is closely related with the next generation matrix (NGM) approach applied in epidemiology (Diekmann et al., 1990; van den Driessche and Watmough, 2002; Heffernan et al., 2005). The elements of the NGM tell us the mean persistence time of a patch or population multiplied by the rate of infection of other patches per unit time, thus the expected number of exported infections during the lifetime of the population infectiousness. The elements of matrix \mathbf{M} , in contrast, tell us the rate at which a patch becomes infected by a specific other patch (rather than infects another patch), multiplied by the expected time that this patch will remain infectious. Feng and DeWoody (2004) also define a slightly different matrix $\tilde{\mathbf{M}}$ that also has λ_M as its leading eigenvalue. Like the NGM, this approach is based upon the analysis of the Jacobian matrix at the trivial solution. So although the approaches correspond in the case of Levins-type models in which $\lambda_I = \lambda_M > 0$, they diverge when the assumptions of these models are broken. That is, the definition of metapopulation capacity threshold allows there to be an occupancy level from which the system shrinks, provided it is growing for at least some p . For example, in the case of a strong metapopulation-level Allee effect, there is a threshold below which the metapopulation shrinks.

In the remainder of this chapter, we use the re-parameterised version of the Levins model (Eqn. 6.3), and the SRLM (Eqn. 6.4) in its usual format. All of the models that we use here are deterministic and their relationship with stochastic models is explored in Section 6.7. We follow Amarasekare (1998) in the choice use the re-parameterised Levins model because the

structural similarity between this version and the logistic model makes it easier to include Allee effects in a way that is consistent with those described in the literature.

6.6 Derivation of SIIS thresholds

In this section we illustrate the utility of viewing models of interest as limiting cases of more general models by deriving the persistence thresholds for spatially implicit and explicit *SIIS* models and using these to find the corresponding thresholds for *SIRS* and *SIS* models. Although we do not explicitly derive these for *SEIS* models, they can be derived in a similar way as for the *SIRS* model. For the spatially-implicit model, we derive the persistence threshold with no Allee effect, a weak Allee effect and a strong Allee effect.

6.6.1 Levins model

We begin by deriving the persistence threshold for the *SIIS* Levins model. The first step is to write the model in the non-standard form. Reformulating the model facilitates the incorporation of an Allee effect in later sections, and in the interests of consistency, we use the nonstandard form throughout. We begin by writing the model in Equation 6.8 in terms of p_{I_1} only. We use two facts: firstly, that $p_{I_1} + p_{I_2} + p_S = 1$; and secondly, because in our model we assume that rates of leaving each of the infectious states are constant, this implies that the time spent in p_{I_2} is proportional to that of p_{I_1} and we can write $p_{I_2} = \frac{r_{I_1}}{r_{I_2}} p_{I_1}$. We obtain

$$\frac{dp_{I_1}}{dt} = (c_{I_1}p_{I_1} + c_{I_2}p_{I_2})(1 - p_{I_1} - p_{I_2}) - r_{I_1}p_{I_1} \quad (6.11)$$

$$= (1 - p_{I_1} - p_{I_2})c_{I_1}p_{I_1} + (1 - p_{I_1} - p_{I_2})c_{I_2}p_{I_2} - r_{I_1}p_{I_1} \quad (6.12)$$

$$= \left(1 - \left[1 + \frac{r_{I_1}}{r_{I_2}}\right] p_{I_1}\right) c_{I_1}p_{I_1} + \left(1 - \left[1 + \frac{r_{I_1}}{r_{I_2}}\right] p_{I_1}\right) c_{I_2} \frac{r_{I_1}}{r_{I_2}} p_{I_1} - r_{I_1}p_{I_1} \quad (6.13)$$

$$= \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) \left(1 - \left[1 + \frac{r_{I_1}}{r_{I_2}}\right] p_{I_1}\right) p_{I_1} - r_{I_1}p_{I_1}. \quad (6.14)$$

We can now re-arrange this model to write the *SIIS* model in the same nonstandard form as for the *SIS* model (i.e. as shown in Equation 6.3) as follows (see Appendix E.2 for the full derivation):

$$\frac{dp_{I_1}}{dt} = \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2} - r_{I_1}\right) p_{I_1} \left[1 - \frac{\left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1}}{1 - \frac{r_{I_1} r_{I_2}}{c_{I_1} r_{I_2} + r_{I_1} c_{I_2}}}\right]. \quad (6.15)$$

As in the *SIS* version, $K = 1 - r/c = 1 - \frac{r_{I_1} r_{I_2}}{c_{I_1} r_{I_2} + r_{I_1} c_{I_2}}$ and since this represents the carrying capacity it is restricted to be positive ($0 < K < 1$), implying that $c > r$. Now,

$\left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1} = p_{I_1} + p_{I_2}$ is the sum of the infectious patches, showing the structural similarity between this expression and the *SIS* version. We can simplify the expression by defining $\beta = c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2} - r_{I_1}$. Now the expression reads

$$\frac{dp_{I_1}}{dt} = \beta \left[1 - \frac{\left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1}}{K} \right] p_{I_1}. \quad (6.16)$$

In order to find the conditions for the existence of a stable non-trivial steady state, we first need to find the non-trivial steady state solution and then evaluate the derivative of the differential equation at this point. Firstly, we observe that at the steady state, either $p_{I_1}^* = 0$ (i.e. the trivial steady state), or

$$1 - \frac{1}{K} \left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1}^* = 0 \iff p_{I_1}^* = \frac{K}{1 + \frac{r_{I_1}}{r_{I_2}}}. \quad (6.17)$$

Differentiating, and evaluating at the nontrivial equilibrium, we obtain

$$D(p_{I_1}^*) = \frac{\partial}{\partial p_{I_1}} \frac{dp_{I_1}}{dt} \Big|_{p_{I_1}^*} = \beta \left[1 - \frac{2}{K} \left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1}^* \right] \quad (6.18)$$

$$= \beta \left[1 - \frac{2}{K} \left(1 + \frac{r_{I_1}}{r_{I_2}}\right) \left(\frac{K}{1 + \frac{r_{I_1}}{r_{I_2}}} \right) \right] \quad (6.19)$$

$$= -\beta. \quad (6.20)$$

For the non-trivial steady state to be stable, we require $D(p_{I_1}^*) < 0$, that is $c_{I_1} + c_{I_2} \frac{r_{I_1}}{r_{I_2}} > r_{I_1}$, thus obtaining the persistence threshold for the *SIIS* model.

We now consider the limits of this threshold for the *SIS* and *SIRS* limiting cases of the model. Recall that we can recover the *SIS* model by allowing $r_{I_2} \rightarrow \infty$. It is straightforward to see that in this case, we obtain the persistence threshold of $c_{I_1} - r_{I_1} > 0$. Similarly, we recover the *SIRS* model by allowing $c_{I_2} \rightarrow 0$, again obtaining a persistence threshold of $c_{I_1} - r_{I_1} > 0$. We can now see that we obtain the same persistence threshold for *SIS* as for *SIRS* because both of these processes have the same effect in the limit on the $c_{I_2} \frac{r_{I_1}}{r_{I_2}}$ term in the expression of the threshold.

6.6.2 Levins model with strong Allee effect

We now consider a modification of the Levins model involving a strong Allee effect. We first apply an Allee threshold $0 < A < K$ to both infectious compartments, such that for $p_{I_1} + p_{I_2} = \left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1} < A$, growth is negative. Following the formulation in Amarasekare

(1998), the dynamics of p_{I_1} can be written

$$\frac{dp_{I_1}}{dt} = \beta \left[1 - \frac{\left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1}}{K} \right] \left[\frac{\left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{I_1}}{K} - \frac{A}{K} \right] p_{I_1} \quad (6.21)$$

$$= \beta \left(1 - \frac{\gamma}{K} p_{I_1} \right) \left(\frac{\gamma}{K} p_{I_1} - \frac{A}{K} \right) p_{I_1} \quad (6.22)$$

where we define the constant $\gamma = 1 + \frac{r_{I_1}}{r_{I_2}}$ to simplify the expression. There are now three steady state solutions, two of which are the same as previously, and a third unstable solution at $p_{I_1}^* = A/\gamma$.

We multiply out to facilitate differentiation, differentiate and evaluate at the nontrivial steady state to obtain

$$D(p_{I_1}^*) = \frac{\partial}{\partial p_{I_1}} \frac{dp_{I_1}}{dt} \Big|_{p_{I_1}^*} = \frac{\beta}{K} \left[2\gamma p_{I_1} - A - \frac{3\gamma^2}{K} p_{I_1}^2 + \frac{2\gamma A}{K} p_{I_1}^2 \right] \quad (6.23)$$

$$= \frac{\beta}{K} \left[2\gamma \left(\frac{K}{1 + \frac{r_{I_1}}{r_{I_2}}} \right) - A - \frac{3\gamma^2}{K} \left(\frac{K}{1 + \frac{r_{I_1}}{r_{I_2}}} \right)^2 + \frac{2\gamma A}{K} \left(\frac{K}{1 + \frac{r_{I_1}}{r_{I_2}}} \right) \right] \quad (6.24)$$

$$= \beta \left(\frac{A}{K} - 1 \right). \quad (6.25)$$

Substituting back in for β and K , simplifying and factoring the expression appropriately, we obtain

$$D(p_{I_1}^*) = A \left(\frac{\beta}{K} \right) - \beta \quad (6.26)$$

$$= A \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2} \right) - \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2} - r_{I_1} \right) \quad (6.27)$$

$$= (A - 1) \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2} \right) + r_{I_1}. \quad (6.28)$$

Now, for the nontrivial solution to be stable, we require $D(p_{I_1}^*) < 0$, giving us a persistence threshold for the model of $(1 - A) \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2} \right) > r_{I_1}$. In other words, the colonisation part of the expression, $c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}$ must compensate for the factor $(1 - A)$, making persistence more difficult.

We can now recover the threshold for the *SIS* and *SIRS* models, with and without a strong Allee effect, by allowing the relevant parameters to tend to 0 or ∞ . We obtain the persistence thresholds for the *SIS* and *SIRS* models as previously as $c_{I_2} \rightarrow 0$ and $r_{I_2} \rightarrow \infty$, in both cases obtaining the persistence threshold in Amarasekare (1998) for the *SIS* model with strong Allee effect. In order to obtain the threshold for the *SIIS* model without an Allee

effect, we allow the Allee threshold, A to tend to zero.

The fact that the threshold is the same for the *SIRS* model with a strong Allee effect as for the *SIS* model may initially seem surprising. However, this is the case because we have applied the Allee threshold to $p_{I_1} + p_{I_2}$. In other words, both infectious compartments contribute to persistence in relation to the Allee threshold, and this is still the case, even when we allow $c_{I_2} \rightarrow 0$, in which the second compartment is now the recovered compartment. This means that we consider recovered patches as contributing to persistence. In order to avoid this, we would need to apply the Allee threshold to the first infectious compartment, p_{I_1} only. The model, when applying the Allee threshold to the first infectious compartment only, can be written

$$\frac{dp_{I_1}}{dt} = \beta \left(1 - \frac{\gamma}{K} p_{I_1}\right) \left(\frac{p_{I_1}}{K} - \frac{A}{K}\right) p_{I_1}. \quad (6.29)$$

Applying the same procedure as above, we obtain the following expression for $D(p_{I_1}^*)$

$$D(p_{I_1}^*) = \frac{\partial}{\partial p_{I_1}} \frac{dp_{I_1}}{dt} \Big|_{p_{I_1}^*} = \beta \left(\frac{A}{K} - \frac{1}{\gamma}\right) \quad (6.30)$$

which differs only in the denominator γ . Rearranging the expression we obtain the threshold condition

$$\left(\frac{r_{I_2}}{r_{I_1} + r_{I_2}} - A\right) \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) > \frac{r_{I_1} r_{I_2}}{r_{I_1} + r_{I_2}}. \quad (6.31)$$

Now, allowing $r_{I_2} \rightarrow \infty$ to obtain $\lim_{r_{I_2} \rightarrow \infty} D(p_{I_1}^*)$, we obtain the *SIS* threshold as $(A - 1)c_{I_1} > r_{I_1}$ as before. However, allowing $c_{I_2} \rightarrow 0$, we obtain the *SIRS* threshold for this model as

$$\left(\frac{r_{I_2}}{r_{I_1} + r_{I_2}} - A\right) c_{I_1} > \frac{r_{I_1} r_{I_2}}{r_{I_1} + r_{I_2}} \quad (6.32)$$

which differs from the original threshold for the *SIRS* model without Allee effect. Allowing $A \rightarrow 0$ recovers the thresholds for the models without an Allee effect, as previously.

6.6.3 Levins model with weak Allee effect

For completeness, we consider the Levins model with a weak Allee effect. Following Tabares and Ferreira (2011), this can be written

$$\frac{dp_{I_1}}{dt} = \beta \left(1 - \frac{p_{I_1} + p_{I_2}}{K} \right) \left(\frac{\varepsilon}{K} (p_{I_1} + p_{I_2}) \right) p_{I_1} \quad (6.33)$$

$$= \beta \left(1 - \frac{\gamma}{K} p_{I_1} \right) \left(\frac{\varepsilon \gamma}{K} p_{I_1} \right) p_{I_1} \quad (6.34)$$

in which $0 < \varepsilon < 1$ is the factor moderating the rate of growth.

In this case, the thresholds can be derived by following the same procedures as in the previous sections. The non-trivial solution is stable provided $\varepsilon \beta K > 0$. Since ε and K are positive by construction, the persistence threshold condition is $\beta > 0$. This threshold is thus the same as in the model without an Allee effect. This is as we would expect: the weak Allee effect influences the rate at which growth occurs in a low-density population, but does not change its sign. The *SIS* and *SIRS* limiting cases of the model also have the same persistence threshold as in the case with no Allee effect.

The *SIIS* model with a strong Allee effect is the most general of the models considered as it subsumes other models of interest as limiting cases and thus can be used to explore the whole family of related models by allowing the relevant parameters to vary.

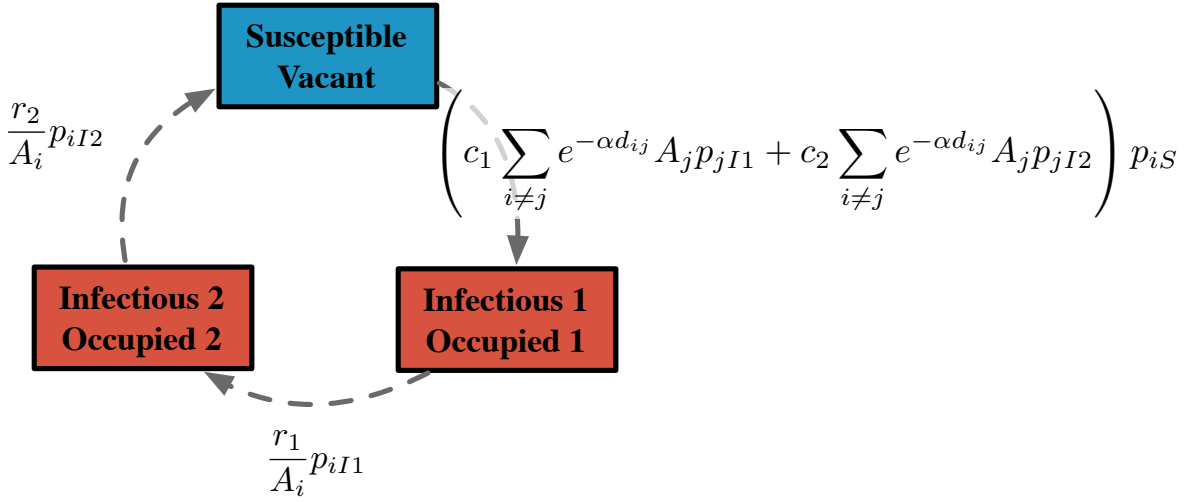
6.6.4 SRLM SIIS

In this section, we derive persistence threshold of the SRLM for *SIIS* dynamics, and show how the threshold under *SIRS* and *SIS* can be recovered as for the Levins model. We do not include an Allee effect, so the invasion threshold is known to be the same as the persistence threshold; however, the threshold for the *SIRS* and *SIIS* models has not been derived previously.

Figure 6.2 shows the *SIIS* dynamics of a single patch under the SRLM model. The relevant equation for the dynamics of the probability p_{iI_1} that patch i is found in first infectious state, can be written

$$\frac{dp_{iI_1}}{dt} = \left(c_{I_1} \sum_{j \neq i} e^{-\alpha d_{ij}} A_j p_{jI_1} + \kappa c_{I_1} \sum_{j \neq i} e^{-\alpha d_{ij}} A_j p_{jI_2} \right) p_{iS} - \frac{r_{I_1}}{A_i} p_{iI_1} \quad (6.35)$$

$$= \frac{c_{I_1}}{A_i} \left[\left(\sum_{j \neq i} e^{-\alpha d_{ij}} A_j A_i p_{jI_1} + \kappa \sum_{j \neq i} e^{-\alpha d_{ij}} A_j A_i p_{jI_2} \right) p_{iS} - \frac{r_{I_1}}{c_{I_1}} p_{iI_1} \right]. \quad (6.36)$$

Figure 6.2: SIIS model. Dynamics and rates for a single patch i .

In this system, A_i represents the area of patch i (or the number of individuals in the population), α is the inverse of the mean migration distance and i and j denote the patches within the system. In a similar way to in the Levins model, the parameters c_{I_1} , c_{I_2} , r_{I_1} , r_{I_2} relate to colonisation and recovery rates from the two infectious compartments; however, the colonisation rates now additionally depend on the distance between patches d_{ij} and the dispersal distance $1/\alpha$ and the area of both the sending patch A_j and receiving patch A_i . Patch recovery rates depend on not only r_{I_1} and r_{I_2} , but also patch area A_i . In order to follow the same threshold derivation process as for the *SIS* model in Feng and DeWoody (2004), we define κ as the ratio of c_{I_1} to c_{I_2} colonisation parameters, such that $c_{I_2} = \kappa c_{I_1}$. We now proceed by first writing the differential equation solely in terms of p_{iI_1} and then solving for the steady state of the system. Furthermore, because patches must transition through all states and the A_i cancel in the expression, as before we also have $p_{iI_2} = \frac{r_{I_1}}{r_{I_2}} p_{iI_1}$ and since state probabilities must sum to one we also have $p_{iS} = 1 - \left(1 + \frac{r_{I_1}}{r_{I_2}}\right) p_{iI_1}$.

Using these relationships, we can write the dynamics of the first infectious state for patch i solely in terms of p_{iI_1} and refactored to write in matrix form, giving (recall that we defined $\gamma = 1 + \frac{r_{I_1}}{r_{I_2}}$)

$$\frac{dp_{iI_1}}{dt} = \frac{c_{I_1}}{A_i} \left[\left(\sum_{j \neq i} e^{-\alpha d_{ij}} A_j A_i p_{jI_1} + \kappa \sum_{j \neq i} e^{-\alpha d_{ij}} A_j A_i \frac{r_{I_1}}{r_{I_2}} p_{jI_1} \right) \left(1 - \left(1 + \frac{r_{I_1}}{r_{I_2}} \right) p_{iI_1} \right) - \frac{r_{I_1}}{c_{I_1}} p_{iI_1} \right] \quad (6.37)$$

$$= \frac{c_{I_1}}{A_i} \left[(\mathbf{M} \mathbf{p}_{I_1})_i \left(1 + \frac{\kappa r_{I_1}}{r_{I_2}} \right) \left(1 - \gamma p_{iI_1} - \frac{r_{I_1}}{c_{I_1}} p_{iI_1} \right) \right]. \quad (6.38)$$

In this form, we can see that at the steady state, we have either $\frac{c_{I_1}}{A_i} p_{i1}^* = 0$ or

$$p_{i1}^* = \frac{(\mathbf{Mp}_{\mathbf{I}_1}^*)_i(\gamma) \left(1 + \frac{\kappa r_{I_1}}{r_{I_2}}\right)}{(\mathbf{Mp}_{\mathbf{I}_1}^*)_i \left(1 + \frac{\kappa r_{I_1}}{r_{I_2}}\right) + \frac{r_{I_1}}{c_{I_1}}}. \quad (6.39)$$

By analogy with the *SIS* case, we can write an iterative map

$$p_{iI_1}^{n+1} = \frac{(\mathbf{Mp}_{\mathbf{I}_1}^n)_i(\gamma) \left(1 + \frac{\kappa r_{I_1}}{r_{I_2}}\right)}{(\mathbf{Mp}_{\mathbf{I}_1}^n)_i \left(1 + \frac{\kappa r_{I_1}}{r_{I_2}}\right) + \frac{r_{I_1}}{c_{I_1}}} \quad (6.40)$$

which converges if and only if the leading eigenvalue λ_M of \mathbf{M}

$$\lambda_M \left(1 + \frac{\kappa r_{I_1}}{r_{I_2}}\right) > \frac{r_{I_1}}{c_{I_1}}. \quad (6.41)$$

It is now interesting to compare this threshold with that from the standard Levins model for the *SIIS* dynamics. Recalling that we chose $\kappa = \frac{c_{I_2}}{c_{I_1}}$, we can now write the threshold as $\lambda_M \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) > r_{I_1}$. In other words, we have a very similar threshold for the SRLM as for the Levins model, with the exception that the SRLM includes the factor of λ_M . We can now calculate the persistence conditions for the *SIRS* and *SIS* models by taking limits of this expression, as we did previously for the Levins model. In both cases, we recover the threshold of $\lambda_M > \frac{r_{I_1}}{c_{I_1}}$.

6.6.5 Comparison of thresholds

A comparison of the thresholds under the different models is shown in Table 6.2. Reading across the rows in the table, it shows that the persistence thresholds under the weak Allee effect are same as those in the model without Allee effect. This occurs because the weak Allee effect decreases the growth rate at small populations sizes but does not alter its sign. The thresholds with a strong Allee effect include a factor of $(1 - A)$; the threshold in these cases forms a constraint for the metapopulation to be growing above this threshold.

Reading down the table, we see that in the case of no Allee effect, our derivations lead us to the same threshold for the *SIS* model as shown in Amarasekare's (1998) earlier work. We have shown that same threshold arises for the *SIRS* model, and for both the *SIS* and *SIRS* with a weak Allee effect. In relation to the model with a strong Allee effect, the derivations based on finding these thresholds as limiting cases of the *SIIS* model allow us to recover the *SIS* persistence threshold found by Amarasekare (1998) but also give us the threshold for the *SIIS* and *SIRS* model. This approach has the advantage of demonstrating that the same

	SIS	SIRS	SIIS
No Allee	$c_{I_1} > r_{I_1}$	$c_{I_1} > r_{I_1}$	$\left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) > r_{I_1}$
Strong (a)	$(1 - A)c_{I_1} > r_{I_1}$	$(1 - A)c_{I_1} > r_{I_1}$	$(1 - A) \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) > r_{I_1}$
Strong (b)	$(1 - A)c_{I_1} > r_{I_1}$	$\left(\frac{r_{I_2}}{r_{I_1} + r_{I_2}} - A\right) c_{I_1} > \frac{r_{I_1} r_{I_2}}{r_{I_1} + r_{I_2}}$	$\left(\frac{r_{I_2}}{r_{I_1} + r_{I_2}} - A\right) \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) > \frac{r_{I_1} r_{I_2}}{r_{I_1} + r_{I_2}}$
Weak Allee	$c_{I_1} > r_{I_1}$	$c_{I_1} > r_{I_1}$	$\left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) > r_{I_1}$
SRLM	$\lambda_M > \frac{r_{I_1}}{c_{I_1}}$	$\lambda_M > \frac{r_{I_1}}{c_{I_1}}$	$\lambda_M \left(c_{I_1} + \frac{r_{I_1}}{r_{I_2}} c_{I_2}\right) > \frac{r_{I_1}}{c_{I_1}}$

Table 6.2: Persistence thresholds for the family of compartmental models. The strong Allee effect has two versions: (a) where the threshold is on $p_{I_1} + p_{I_2}$ and (b) where is it on p_{I_1} only.

threshold arises for the *SIS* as for the *SIRS* models because both $r_{I_2} \rightarrow \infty$ and $c_{I_2} \rightarrow 0$ have the same effect on the $c_{I_2} \frac{r_{I_1}}{r_{I_2}}$ term in the expression, which tends to zero in both cases. Comparisons between the Levins and SRLM models show that for the latter there is the additional factor of λ_M . This can be thought of as accounting for factors such as patch sizes and connectivity that influence extinction rates within populations and colonisation between them in this more realistic model.

One implication of the findings for work on persistence is that models with a strong Allee effect may be more appropriate for considering persistence in stochastic systems. Since in the cases with no Allee effect and a weak Allee effect, the equilibrium proportion of infectious individuals can become arbitrarily small even above the persistence threshold, the threshold may not be particularly informative about real populations in which stochastic extinction is likely to occur rapidly in such cases. In this case, the model with a strong Allee effect, in which a minimum (non-zero) population threshold can be defined, may provide more sensible predictions about persistence thresholds in stochastic settings.

6.7 Discussion and conclusions

In this chapter, we presented a new overarching framework for considering the relationships between various kinds of models. We began by discussing the relationship between compartmental models used in epidemiology and ecology and their possible interpretations in the two contexts, at the individual and metapopulation scales. Although some of these connections have been noted, the discussion presented here is much more comprehensive. It is helpful

because it can be used to guide the interpretation of findings from one application area into the other when the models used are mathematically identical. The use of compartmental models in ecology is generally less standardised than in epidemiology, and this discussion may suggest new opportunities to employ the findings from epidemiological theory. It also makes more explicit the connection between compartmental models at different scales.

As part of this discussion, we clarified the definitions of metapopulation capacity in earlier work, as well as making more explicit the connections between the definition of metapopulation capacity, the intuition behind its interpretation in relation to persistence, and its calculation. We also clarified the relationship between the ecological concepts of metapopulation persistence capacity λ_M and metapopulation invasion capacity, and the epidemiological concept of the basic reproduction number R_0 as viewed from the next generation matrix perspective.

In a second aspect of making explicit the connections between model types, we showed how a set of models that are typically considered separately - namely the *SIS*, *SEIS* and *SIRS* models - can be considered as a family of models through their connection with the more general *SIIS* model. Perhaps because of the deliberate discretisation of states in compartmental models, the connections between these models into a family have typically gone unacknowledged. Yet we believe that viewing the connections explicitly can have multiple benefits for ecological and epidemiological theory.

In order to demonstrate one potential use of the connection between models within this family, we considered persistence thresholds as an example. We described how the standard approach for establishing the persistence threshold for the simple *SIRS* model makes use of mathematical facts that not only obscure the reasons for which the persistence threshold in the *SIRS* model is the same as that for the *SIS* model, but also fail in the case of a strong Allee effect. We add to the family of Levins models versions with strong and weak Allee effects, and the Spatially Realistic Levins Model (SRLM).

We then provided derivations of persistence thresholds that make use of the relationship between models in the family. First considering the Levins model for *SIIS* dynamics with no Allee effect, a strong Allee effect and a weak Allee effect, we derived the respective persistence thresholds. To our knowledge, these derivations are new for the *SIIS* model. We then allowed the relevant parameters to tend to the appropriate limits to recover the *SIS* and *SIRS* model thresholds, showing that in the case of the *SIS* model, these correspond to earlier results. Because of the mathematical equivalence between the Levins and logistic growth models, these thresholds also apply to the logistic growth model. We derived the persistence threshold for the SRLM under *SIIS* dynamics (with no Allee effect) and showed how this can be used to extract the persistence thresholds for the *SIS* and *SIRS*. The derivation for the *SIIS* and *SIRS* models is new, and we showed that in the *SIS* case,

the threshold corresponded to earlier results. We commented on the connection between the persistence threshold for the Levins and SRLM models. For the thresholds that have been derived previously for *SIS* models, we note that the approach taken here, deriving the thresholds for a more general model and deriving thresholds for simpler models from it, is new.

We have shown that by considering specific models as limiting cases of more general models, we can gain useful information about the relationships between persistence thresholds within the family. Specifically, we showed using this approach why the persistence threshold in Levins models is the same for *SIS* as it is for *SIRS* models: the two processes by which the *SIIS* tends to the *SIS* and *SIRS* models have the same effect on the terms of the expression of this threshold. Further, we derived the persistence threshold of the Levins model under both weak and strong Allee effects, showing that for the models in the family, the thresholds differ between those with a weak and strong Allee effect, and for some models, also depend on the application of the Allee threshold to one or both of the infectious compartments. Although we did not provide derivations for the *SEIS* model, obtaining the thresholds for this model from the *SIIS* model follows a similar approach to the *SIRS* model, but where $c_{I_1} \rightarrow \infty$. To make this more meaningful, instead of writing the differential equation for p_{I_1} , it is probably clearer to write the differential equation for p_{I_2} . In the case of the SRLM, we derived the persistence thresholds for the model without an Allee effect under the different dynamics.

The approach taken here allowed us to explore connections between the concept of interest, persistence thresholds, within a specific family of models. We note that although deriving the thresholds for the *SIIS* model was more complex than doing so for either of the individual models *SIS* or *SIRS* (or *SEIS*), it was more efficient than doing so for each of the models independently. The approach could be extended to other concepts of interest such as invasion thresholds or equilibrium values, or more generally to other families of models. In addition to shedding light on the connection between thresholds derived in the chapter, the relationship between thresholds for the Levins model and SRLM suggest hypotheses regarding the persistence threshold for the version of this model with Allee effects that could be explored in future work. In other words, the approach leads us not only to new questions but also new suggestions regarding their answers. In addition, the question arises of how persistence thresholds are affected by Allee effects that apply only to one of the infectious components, or that affect only colonisation or extinction but not both.

Another possible application of the approach is that of providing guidance about the most appropriate model structure for a particular application. We noted earlier that choices between different models within the family are rarely well motivated in the literature. In his review of persistence thresholds, Roberts (2007) treats different disease models in separate sections: the *SIRS* model is considered apart from a model with a carrier class in which some individ-

uals recover from acute infection into a class from which they are able to transmit infection at a reduced rate. Consideration of these different models highlights the importance of making an informed choice of the disease classes to include in compartmental models and the relationship between these classes. For example, the carrier model demonstrates that the inclusion of a carrier class can lead to different conclusions about disease persistence and suggests that decisions about which disease classes to use should be made carefully when considering persistence thresholds. For example, we may wish to ask whether it is important to allow recovered individuals to continue to transmit disease by changing from an *SIRS* to an *SIIS* model in which the two *I* compartments have different transmission rates, and how this changes persistence. Analogous questions can be asked when considering persistence thresholds in ecology, such as that of the extent to which persistence affected by allowing regenerating habitat to be re-colonised at low rates. The connections between models can therefore be used to explore the sensitivity of findings to the particular choice of model. Our framework allows us to quantify the effects of selecting, for example, an *SIRS* model as opposed to an *SIIS* model in which the second infectious compartment demonstrates less infectivity than the first.

Mathematical and other challenges remain in order for the benefits of the approach to be fully realised. In relation to the SRLM and *SIIS* version thereof, the assumptions made in the above sections imply that this threshold can only be derived in this way for a particular family of models. Specifically, the derivation can be extended to models with an arbitrary number of infectious states, provided that the time spent in each state (or the rate at which patches leave the state r_n) is constant, and thus can be written as proportional to the time spent in the first infectious state. The rate could, nonetheless, be a different function of A_i such as $\frac{1}{A_i^n}$. A similar situation is true for the colonisation parameters, which must again be constant. In order to employ the derivation using matrices, the colonisation rate of a patch must be a linear function of the colonisation potential of the donating patches. Incorporation of an Allee effect therefore remains for future work. A threshold exists even if these conditions are not met in the proportionality of the rates; however, it is no longer possible in these cases to write a matrix M and the threshold must then be found numerically.

In relation to the implications of these findings for persistence, an important consideration is that of the extent to which persistence threshold derived for deterministic models such as these apply to stochastic systems. Their application to stochastic models suffers from at least two problems. Firstly, one can be above the persistence threshold in a deterministic model and yet equilibrium values can be arbitrarily close to zero. This is particularly problematic in the case of the *SIRS* model in which equilibrium values are reduced by the existence of the *R* compartment, yet the persistence threshold remains unchanged. The question of the effect of a recovered stage (or, symmetrically, an exposed stage) on real-world persistence remains unclear. The analysis of the model without an Allee effect suggests that this will depress

the number of observed cases, but should not affect persistence. On the other hand, one might argue that depressing the number of infectious agents should make populations more susceptible to stochastic fluctuation and lead to the prediction that even if these systems can persist in a deterministic model, they should show a strong tendency to rapid extinction in a stochastic model. It may be possible to investigate which of these explanations corresponds to real-world systems by making comparisons between the model with an Allee effect that applies to only one of the infectious stages and one that applies to both. More generally, the use of an Allee threshold should assist with the problem of arbitrarily small equilibrium population sizes. In order to account for arbitrarily small population sizes in deterministic models, one could use a model with a strong Allee effect, setting the Allee threshold to a population size that is expected to be sufficiently large to avoid rapid stochastic extinction.

The second problem in relation to applications in a stochastic setting is that of the introduction of additional infectious states. The direct stochastic equivalent of the deterministic model assumes exponentially distributed waiting times until the transition from one state to another. However, the use of multiple exponentially distributed waiting times together for a single process gives rise to a gamma distribution. It remains for future work to establish the extent to which this affects persistence in relation to the thresholds calculated here.

Chapter 7

Discussion

The main aims of this thesis were to develop our understanding of biological persistence, and to explore the role that modelling can play in this development. This was achieved through four studies that explored the effects of (1) modelling decisions that relate to the relationship between the processes and structures of the system and those embodied in the model and (2) measurement decisions about how to capture and characterise persistence. The main modelling decisions compared were whether time is modelled discretely or continuously, different ways of characterising the contribution of a patch to metapopulation persistence, different ways of representing the long-run distribution of system states, and the effect of changing the states through which a patch transitions. Persistence was characterised in different chapters as continued existence until a pre-defined target time, mean time to extinction, patch occupancy probability distributions in the long run, and as the existence of a stable non-trivial equilibrium in a deterministic model. We now provide an overview of each of the chapters.

7.1 Overview

Chapter 3 consists of a stochastic simulation study using a cellular automaton approach, in which organisms live on a regular lattice, and explored the effects on persistence and inter-specific competition of the decision to model time as continuous or discrete. The model was simulated stochastically as a cellular automaton in continuous and discrete time, considering discrete time as an approximation to the ‘true’ model in continuous time. We used a spatial version of the Gillespie Algorithm for the continuous time approach and a range of update rules in the discrete time approach. Population dynamics were modelled according to a lattice logistic growth model, selected to facilitate comparison with other work on the effect of update schemes in the literature, especially Ruxton and Saravia (1998). We used initial population sizes of 10, 100 and 1000 on a 100×100 grid and as a simple persistence measure, we

use a time horizon of 1000 time units, with a census taken at this point. We found quantitative and qualitative differences in persistence between populations simulated in discrete and continuous time, and that qualitative conclusions about species dominance can depend on the decision to model time in one of these two ways. We argued that it is important to consider the decision about how to model time when asking ecological questions of this type. Further, we noted that the discrete time approach also entails additional modelling decisions about the ordering of births and deaths, and an appropriate conversion mechanism between continuous and discrete time which may also affect persistence conclusions. We also showed, using a sample run, that the time-horizon approach to measuring persistence can be misleading in the sense that extrapolations to different time horizons can be qualitatively different depending on the rate of population growth or decline; conclusions drawn from this measure are strongly affected by the horizon selected, and it is therefore rather unsatisfactory as a way of characterising persistence. This chapter has previously been published as Mancy et al. (2013) and provides a novel contribution to the literature by highlighting the importance of the decision to model processes in discrete or continuous time, and specifically the difficulties of approximating continuous spatial processes by discrete time simulations, especially in systems with more than one species.

From the perspective of computational biology, the chapter describes a comparison between existing simulation algorithms used in contemporary computational biology, including ecology and epidemiology¹. Our aim was thus not to develop new algorithms or improve upon existing algorithms but to conduct a computational experiment to support us in providing guidance on modelling decisions that must be made when constructing simulation platforms for studying biological systems. Our main interest was in the correspondence between the output of these algorithms, rather than on their algorithmic complexity or speed of execution. From an experimental algorithmics perspective, the key question of the chapter is analytic, focusing on the conditions under which a collection of algorithms in which time was discretised, are able to provide acceptable accuracy as approximations to a benchmark algorithm that assumed continuous time. Another way of thinking about this problem, in the language of Polack (2014), is that the question is one of the validation of simulation models (in discrete time) in comparison with a simulation model (in continuous time) that is already accepted as valid. This work contributes to the computational biology literature by demonstrating a methodological point: choosing a discrete-time formalism in construction the platform model, as opposed to a continuous-time one, can have important implications for the biological conclusions drawn.

¹The algorithms used in this literature are rarely made fully explicit, making it difficult to assess whether the fine-level details of the implementation (e.g. the data structures employed) of our algorithms are new; however, distinguishing between abstract algorithms (as typically expressed in pseudocode) and coded implementations (McGeoch, 2007), computational approaches that with the same or very similar abstract descriptions exist in the literature.

Chapters 4 and 5 focus on persistence using the notion of the quasi-stationary distribution. The real-world problems from ecology and epidemiology that provided our initial motivation for investigating the QSD were whether it is possible to predict (1) the expected persistence time and (2) the effect on expected persistence time of the removal of a particular patch, of an ecological or epidemiological system that has already persisted for a long time and thus might reasonably be assumed to be in the quasi-stationary regime. The expected persistence time from quasi-stationarity can be found from the QSD, while the contribution of a patch to the QSD (and thus to persistence time) has not been investigated previously. From a computational biology perspective, the main computational problem posed by these questions is that the memory requirements for the representation of the QSD (and thus for its computation) make the problem intractable, even for modestly-sized inputs. For the model we consider, the size of the algorithm output (i.e. the QSD) is determined by the input in the form of the number of patches n ; specifically, the space requirements of representing the output scale as $O(2^n)$. Because the issue of dimensionality is seen in the size of the required output, it is inherent to the problem itself (although it can be worsened by inefficient algorithms), and algorithmic developments need to focus on useful computable approximations to the QSD. This work in these chapters therefore addresses the classical question underpinning much of computing science of ‘what can be (efficiently) automated’ (Denning et al., 1989) as applied to the computation of biologically useful approximations to the QSD and related quantities, for a particular class of ecological models. It contributes to the computational biology literature by providing an analysis of the accuracy of a range of algorithms for approximating the QSD.

From a biological perspective, Chapter 4 focuses on the question of how to model the contribution of a patch to metapopulation persistence. The theoretical model on which the study is based is the spatially realistic Levins model (SRLM), that models a metapopulation as a finite collection of patches that can either be occupied or vacant. We used a stochastic version of the model in which patches are colonised and go extinct according to characteristics such as their area and spatial location. As our measure of persistence, we employed the mean time to extinction from quasi-stationarity. We used a standard numerical approach to calculate the mean time to extinction from quasi-stationarity for randomly-generated landscapes under a range of parameter values. We calculated the difference between the mean time to extinction for the original landscapes, and from each landscape with a single patch removed, in order to obtain a measure of the contribution of the individual patch to metapopulation persistence. We then tested a range of measures of patch value that have been proposed in the literature. Our findings show that under the standard SRLM, the square of the proportion of time that a patch is occupied in the QSD provides a good approximation of the proportional contribution of the patch to mean time to metapopulation extinction, clearly out-performing the patch value measures proposed by Ovaskainen and Hanski (2001); Ovaskainen (2003).

Testing with a more general version of the SRLM in which patch characteristics were raised to arbitrary powers to give their effect on emigration, immigration and extinction showed that the findings generalise relatively well in most parts of the parameter space evaluated, although for some parameter sets, it was preferable to raise the occupancy probability to a different power. The chapter contributes to the theoretical literature by identifying this connection and raising further questions about the exact relationship between patch occupancy probability and patch value. Perhaps more importantly, this work has the potential for significant contributions to interventions and real-world applications to either protect or eliminate populations since it suggests a relatively robust estimate of the contribution of a patch or population based on only readily available data.

From a computational point of view, our aim was to find a representation of the contribution of patch that avoided the problem of the space complexity by circumventing the necessity to compute the QSD. We used an experimental computing science approach to compare different candidate measures of patch contributions found in the literature, as well as other measures that employed these. Working with computationally tractable systems for which the true (patch) contributions could be established, our approach involved a search for regularities in the relationship between these candidate measures and true contributions. The experimental approach taken was iterative: initial candidate measures were derived from the literature and tested for their relationship with true contributions; based on these findings, adjusted candidate measures were developed and again tested. Although we tested candidate measures that would have required complex model fitting before application to real-world problems (thus requiring computational science in their application), we found that the best of the candidate measures could be easily calculated directly from real-world data. Interestingly, our computational biology study thus allowed us to suggest a heuristic that, at least in principle, can be used in applied biological contexts without recourse to further use of computational science methods. The study contributes to the computational biology literature by showing the relationship between the ‘exact’ and the heuristic in the context of an ecological model. More generally, it provides a demonstration of the use of computational biology techniques to identify simple relationships in models. These relationships, in turn, can be thought of as hypotheses for empirical, real-world testing, thus demonstrating the role of computational biology in the generation of scientific hypotheses. In addition, the study provides a demonstration of the use of computational biology techniques to uncover simple principles that can act as heuristics to guide interventions in the systems represented, that no longer require complex modelling efforts.

This work on patch contributions demonstrates the potential of the QSD to contribute to our understandings of persistence; however, it has typically remained under-utilised in practical applications, probably in part because of the difficulties associated with the exponential growth in the state space for heterogeneous models.

In Chapter 5, we focus on the problem of simulating biologically useful approximations to the full QSD. We take as our starting point an existing computational science method for simulating the quasi-stationary distribution that has been described previously in the physics literature. In this literature, the simulation algorithm described is applied to two models of interest in the physics literature (de Oliveira and Dickman, 2005; De Oliveira and Dickman, 2006). Our aim in the work described here was to further develop this algorithm to simulate approximations to the quasi-stationary distribution of an ecological model that incorporates higher levels of heterogeneity than those in the physics literature, and to do so in ways that are useful to practising ecologists. This chapter therefore takes an engineering approach to algorithm development, with experiments conducted to compare candidate approximations. Although de Oliveira and Dickman (2005) describe their work as providing an algorithm for computing the quasi-stationary distribution, the distribution tracked in the simulations provided as illustrations of the approach is not that of the full system, but a marginal distribution: instead of computing the probability distribution of site occupancy patterns, they compute the probability distribution of the number of occupied patches. In other words, they reduce the dimensionality of the problem by considering site occupancy patterns as equivalent. The introduction of heterogeneity in the ecological model in our work meant that it was no longer appropriate to consider these patterns as equivalent, and the marginal distribution thus fails to capture the information of interest.

In this chapter, we therefore considered computationally tractable candidate compressions of the QSD. From an experimental algorithmics perspective, the problem is one of algorithm design, and focused on the development of algorithms for approximating the QSD that have lower space complexity. Our study shows that for examining common system states, a clustering approach provides the best approximation among those tested. For examining time to extinction, while also permitting an analysis of near-extinction states, we suggest the use of a simpler system named ‘independent patches by occupancy’. From a biological perspective, the findings reported provide a practical tool to guide the modelling decisions of researchers interested in the QSD for these kinds of systems. From a computational biology perspective, the work in this chapter contributes to the literature through the development of algorithms for the simulation of the QSD for a model used in the ecological literature, and through guidance on how to choose between these algorithms for particular applications, as well as by raising engineering questions related to the possible ways to improve upon the compression models considered. We note that the model is very closely linked with other lattice and network models, and we expect that these algorithms (or close variants of them) should be more widely useful.

In Chapter 6, we considered a range of deterministic metapopulation models as a way to investigate the extent to which these can provide useful information about the persistence of real-world systems. Specifically, we considered the notion of persistence thresholds for

the Levins model and SRLM². Persistence was operationalised as the existence of a stable nontrivial equilibrium in the model. As part of our discussion, we clarified the definitions of metapopulation capacity of the SRLM provided in earlier work, as well as making more explicit the connections between the definition of metapopulation capacity, the intuition behind its interpretation in relation to persistence, and its calculation. We also clarified the relationship between the ecological concepts of metapopulation persistence capacity and metapopulation invasion capacity, and the epidemiological concept of the basic reproduction number R_0 as viewed from the next generation matrix perspective.

In relation to persistence, we investigated the effect on persistence thresholds of modelling decisions regarding the states through which a patch progresses. Using epidemiological terminology to describe these processes, we extended the original presence-absence or susceptible-infectious-susceptible (*SIS*) model to capture an additional state. We introduced a model in which each element transitions through a second infectious state (*SIIS*), that rarely appears in the literature, as a way to link models with two and three states. This model subsumes the model with an exposed state (*SEIS*) and that with a recovered state (*SIRS*), as well as the *SIS* in the limit. This framework allowed us to derive new persistence thresholds and obtain existing ones by first deriving these for the *SIIS* model and then allowing key parameters to tend to limits that allowed us to recover more specific models. The ability to do this becomes important for making modelling decisions when we are not certain, for example, about whether individuals or patches that are recovered are still infectious, even if at a lower rate (and thus whether an *SIIS* or *SIRS* model is more appropriate). It also allows us to explore the relationship between persistence thresholds for models with different numbers of states and those with and without Allee effects that cause populations to go extinct when occupancy is low. These suggest that including an Allee threshold in models may make them more appropriate for the application of deterministic persistence thresholds to stochastic models. The main novel contribution of this chapter rests on the framework that underpins the derivation of thresholds for linked models that can be considered as limiting cases of a more general model. This approach could be applied more widely, such as to guide modelling decisions about which states to include in compartmental models. A number of the persistence thresholds derived are also new. The work in this chapter arises from considerations developed in connection with our contributions to the development of the notion of reservoir capacity in Viana et al. (2014), co-authored during the period of the PhD.

We now discuss the implications of these studies for future work on biological persistence and the role modelling can play in that endeavour.

²The Levins model is a simpler version of the SRLM in which there are infinite, identical patches, and is mathematically equivalent to the logistic growth model, a spatial version of which was employed in Chapter 3

7.2 Implications for future work

7.2.1 Model types and persistence

The investigations carried out in this thesis have highlighted the distinctions between deterministic and stochastic models for studying persistence and the tradeoffs between modelling decisions. Specifically, deterministic models fail to capture the fact that extinction is eventually certain, while relevant stochastic concepts such as the quasi-stationary distribution suffer from the difficulties of a very large state space. It would therefore be helpful to explore ways to improve our understanding of both deterministic and stochastic systems to overcome these problems.

Due to the lower complexity of deterministic models, to better understand the persistence of biological populations, it would be valuable to learn more about how to map between deterministic and stochastic models, since the most obvious mapping is problematic because deterministic models predict persistence even for arbitrarily small populations. One possibility is to consider including a strong Allee effect, the threshold for which might depend on population size, in considering persistence of stochastic populations. It would be helpful to know, for example, how to set the threshold of the Allee effect to obtain information about mean time to extinction. An alternative approach might consist of using complementary information, perhaps by combining information on the persistence threshold and stability of the steady state solution. This may be appropriate since higher persistence in stochastic models might be expected to be associated with situations in which the nontrivial steady state is both relatively large (far away from extinction) and one that a system returns to strongly and rapidly following perturbations.

In relation to stochastic approaches, work to improve the tractability of the quasi-stationary distribution would be valuable. Work on tractable simulation approaches remains very underdeveloped, yet could provide practical solutions to problems of the complexity of the state space. We have shown that it is possible to provide relatively good estimates of the QSD for heterogeneous systems of intermediate size through simulation that employs compressed versions of the QSD, and it should be possible to improve the schemes presented here. For example, it seems likely that an approach that uses the clustering algorithm for clustering states, after conditioning on the number of occupied patches. Other developments might be found in the way in which QSDs are compared. In our work, we compared the probability distribution of discrete states; however, it is reasonable to consider some of these states as more similar than others (e.g. a state in which all patches of a relatively large system are occupied and one in which all except one are occupied are more similar to one another than either is to one in which a single patch is occupied). It would be very valuable to develop more sophisticated measures of the distance between QSD representations that take account

of this similarity, and to explore the accuracy of existing and new algorithms for simulating the QSD with respect to these. A second alternative is to establish the statistical relationships between aspects of interest of the QSD, such that the information that we require can be obtained directly. For example, here we have shown simple predictors of the contribution of a patch to metapopulation extinction times and size. The extent to which these relationships hold in different models, the mathematical reasons for the identified relationships (i.e. how to formalise and understand patch values), and the exploration of other possible relationships of interest would therefore represent a valuable endeavour.

The applicability of the QSD to some of the questions investigated here should also be considered in more detail. For example, once a system has been perturbed (e.g. by patch removal), the distribution of states immediately following perturbation may show little resemblance to the QSD into which the system eventually relaxes, and the time taken for this to occur is important if we are to make predictions about likely future states following perturbation or to apply several interventions in sequence. The question of transients therefore becomes important (Hastings, 2004). It also remains unclear how well a system is represented by the QSD in the case of systems below the critical threshold. The ratio of expectations (Artalejo and Lopez-Herrero, 2010), which is the distribution of system states conditioned on non-extinction from a given starting state better captures this situation; however for unknown starting states, the ratio of expectations must be calculated for all initial conditions, and the state space is thus even larger than that of the QSD.

7.2.2 Questions arising from possible applications

The study in this thesis that appears to have the strongest policy implications is the chapter on patch values. This work provides a very simple heuristic that could be used to guide interventions; however, the applicability of this heuristic to the real world remains to be demonstrated, and at present, it would be wise to treat it with considerable caution. Firstly, in relation to real world systems and policy questions, it would be very valuable to consider in more detail the extent to which particular systems map to the models discussed in this thesis, both in the context of ecological and epidemiological systems, since the patch value heuristics proposed here work better under some model assumptions than others (e.g. when the importance of the characteristics of the ‘parent’ or sending patch are similar or less strong than those of the receiving patch). In attempting to map these systems to those in the real world, it would be helpful to consider exactly what is meant by each of the states, and whether, for example, we should choose to consider that a patch is occupied from the point of invasion or from the point at which a population becomes ‘established’ there (the latter seems to have a better mathematical foundation since we know that extinction times are exponentially distributed once the QSD is reached). At least a partial answer to this question

could be found through a simulation approach using an individual-based model.

The extent to which the findings hold in systems with individuality is also an important question that needs to be addressed, although we note that for similar systems, Ovaskainen and Hanski (2004) find good correspondence. Additional difficulties arise in measuring the relevant information to parameterise the model. For example, patches are often poorly delimited in real systems. As a result, it would be extremely valuable to explore the extent to which choosing the boundary of patches matters to the heuristics described here. Exploring this by considering different patch groupings may be a good way to use the existing framework to explore these questions. In addition, few systems are truly closed (there are typically imports of pathogens or organisms from external sources), and it would be useful to understand the extent to which these can be incorporated into the work described here.

Extensions to the current work also include the consideration of additional patch states (e.g. exposed, recovered) and whether the patch contribution changes as a function of these. It would also be useful to extend to multi-species models. A further extension would be to explore the effect of non-exponentially distributed waiting times in simulating transitions between states.

7.3 Concluding remarks

In summary, in this thesis, we have explored a range of theoretical and practical issues pertaining to biological persistence through four studies focusing on different measures of persistence and the effect of modelling decisions on the conclusions that can be drawn from them. From the broad perspective of understanding persistence, the main contributions of the thesis are in highlighting the extent to which different measures of persistence - including those defined by time horizons, mean time to extinction from quasi-stationarity, and the existence of a stable nontrivial deterministic steady state - can lead to different conclusions about biological persistence. In relation to modelling decisions, the contribution is in the demonstration of the importance of the choice of discrete or continuous time, of ways of modelling the contribution of a patch, of particular assumptions underpinning compressions in simulating the QSD, and of which compartments are included in compartmental models with multiple states and the functional forms governing how metapopulation growth relates to occupancy (with or without Allee effects). The sections above outline priorities for future research that we hope will contribute to the further development of theory and to conservation biology and infectious disease control.

Appendix A

Chapter 1

A.1 Great principles of computing science

In this appendix, we describe how we have employed the principles of computing science explained in Denning (2003).

Table A.1: Application of principles of design

Principle	My application
Simplicity Forms of abstraction and structure that overcome the apparent complexity of applications	Use of object-oriented code for the main simulations, use of style sheets for the figures in R/ggplot
Performance predicting throughput, response times, bottlenecks, capacity planning	Code written to manage memory (e.g. via simulation to avoid reliance on storing full transition matrix in Chapter 5)
Reliability Redundancy, recovery, checkpoints, integrity, system trust	The simulation platform was not mission critical and therefore reliability was a lesser concern than for many commercial applications. Nonetheless, code was developed such that simulation output was stored separately from error messages. Error files were produced by the architecture managing the different concurrent simulations (cluster), as well as directly by the simulator code, the latter of which contained Java errors as well as checkpoint information. These allowed us to conduct searches within output files using the ‘grep’ command and keywords such ‘error’ to identify if there had been any problems during the running of the simulations.
Evolvability Adapting to changes in function and scale	Many versions of the code were written, and the final structure was chosen so that the most obvious new functionality could be added in a straightforward manner. For example, in Chapter 5, the QSD representations were stored as different instantiations of an abstract class, so that new compressions could be added without modifying other code.
Security Access control, secrecy, privacy, authentication, integrity, safety	The issue of security was largely irrelevant for this work; instead, a commitment to code sharing and documentation was preferred.

Table A.2: Application of principles of mechanics

Principle	My application
Computation What can be computed; limits of computing	Consideration of the limits of computation in calculating the QSD.
Communication Sending messages from one point to another	Denning includes file compression in this category, and although communication was not our primary concern in using compression in Chapter 5, this fits in this category.
Coordination Multiple entities cooperating towards a single result	Different runs of the simulator under different parameter values used to answer scientific questions.
Automation Meaning and limits of automation, performing cognitive tasks by computer	For example, a machine learning approach employed in Chapter 5 to cluster system states.
Recollection Storing and retrieving information	Output files were stored by the name of the code version and the experiment, and a combination of parameter values, so that their contents could be easily identified.

Table A.3: Application of practices

Practice	My application
Programming Computing professionals must be multilingual, facile with numerous programming languages, each attuned to its own strategies for solving problems	<p>Four main computer languages were used in the work described in this thesis.</p> <ol style="list-style-type: none"> 1. The main code of simulations was developed in Java. This was appropriate because these were liable to updating and the modular structure of object-oriented programming made this a sensible approach. In addition, the relative speed and open source nature of the language suited the application and scientific computation areas. 2. Matlab was used as appropriate for rapid prototyping, and also for numerical solutions to differential equations and matrix problems. 3. R was used as this language is appropriate for final statistical processing of simulation output and construction of figures. 4. Unix / Linux shell scripting was used to coordinate the simulations. The cluster on which most of the work was conducted required this language.
Engineering systems To produce a valuable benefit for users; also an engineering component concerned with the modules, abstractions, revisions, design decisions and risks; and an operations component configuration, management and maintenance of the system	<ol style="list-style-type: none"> 1. Benefit for users: The main users of the simulation platforms described in the thesis are the developers and collaborators, as well as readers of the research publications resulting from the work. The systems are successful insofar as this has been possible and that they can be employed for future research. 2. Engineering: The systems have been designed in ways that allows their reuse. 3. Operations: The simulation platforms have been maintained and will be made available online.

Practice	My application
<p>Modelling and validation Building models of systems to make predictions about their behaviour under various conditions; and designing experiments to validate algorithms and systems</p>	<p>In scientific computing, computing practitioners construct models not of computer systems, but of natural systems. The main aim of the work in this thesis was to build models that make predictions about the behaviour of these systems under a range of conditions. To do this, experiments were designed and executed to validate these models. In most parts of the thesis, approximations are compared to a baseline model. One aspect that is not reported in detail is that when developing the simulators themselves (as opposed to the full simulation platform), two copies of the code were developed and compared. In Chapter 3, two versions were developed, each in Java, but by two independent researchers. In Chapter 4 and Chapter 5, one version of the simulator to produce the full QSD was developed in Matlab by constructing the full $2^n \times 2^n$ transition rates matrix for the system, and a second version was developed in Java, with transition rates calculated on the fly as a function of the current system state. The existence of two independent versions of each of the simulation platforms allowed us to identify bugs much more easily than if a single system had been developed.</p>
<p>Innovating Exercising leadership to design and bring about lasting changes to the ways that groups and communities operate. Innovators watch for and analyse opportunities, listen to customers and deliver promised results</p>	<p>Customers, in the context of scientific computation, are typically practising domain scientists. The three computational chapters of this thesis each constitute a response to problems raised by scientists, mostly in informal discussions I have had with them. The focus on methodological questions of interest to practising scientists should help to support lasting changes in practice among this group.</p>
<p>Applying Working with practitioners in application domains to produce computing systems that support their work</p>	<p>The practitioners in the application domains of the work in this thesis are mostly scientists in ecology and epidemiology. The work in this thesis has been conducted in discussion with members of this group.</p>

Appendix B

Chapter 3

B.1 Algorithms

The algorithms *Gill*, *RF_d2S*, *RF_d2M* and *RR1S* are explained below using the notation in Table B.1. The algorithms are described for multiple species, where the species is denoted by an identifier in the form of the subscript i . All algorithms take the following arguments (additional arguments are described in the text of each algorithm): the maximum time t_{max} and habitat occupancy A , a data structure in which the index represents site and the entry is either empty or holds the value of i for the species living at the site. Superscripts r and p represent rates and probabilities respectively.

B.1.1 *Gill*

The *Gill* algorithm (space constrained Gillespie with multiple species) is presented in Algorithm B.1. In addition to t_{max} and A , it takes as arguments a set \mathcal{S} of species, two vectors containing the *per capita* birth and death rate for each species b^r and d^r , and a vector N containing the total number of organisms of each species currently inhabiting the grid.

The algorithm iterates until the maximum time is reached (outer `while` loop) and has three main sections: lines 5-9 compute the relevant population level rates, lines 10-23 compute the next event and lines 24-25 compute the elapsed time. In lines 5-9 the population level birth and death rates (B^r and D^r) are computed for each species by multiplying per capita rates by the number of organisms of that species, and the total event rate λ is computed as the sum of all event rates (over all species). In lines 10-23, species are considered in turn and the probability $\delta \in [0, 1)$ that the next event is a death of the current species is computed by normalising the species death rate; the probability $\beta \in [0, 1)$ that the next event is a birth of the current species is calculated analogously. The algorithm then compares δ and β to the

Table B.1: Symbols used in algorithms

Symbol	Description	Algorithms
i	Species identifier	All
t, t_{max}	Time, maximum time	All
A, B	Habitat occupancy; index gives site and entry contains species identifier i	All
b_i^r, d_i^r	<i>Per capita</i> birth and death rate for species i	<i>Gill</i>
B_i^r, D_i^r	Total population birth and death rate summed over individuals of species i	<i>Gill</i>
N_i, N	Number of individuals of species i , total number of organisms	<i>Gill</i>
τ	Time until the next event under <i>Gill</i>	<i>Gill</i>
λ	Total event rate	<i>Gill</i>
\mathcal{S}	Set of species	<i>Gill</i>
β, δ	Probability that the next event is a birth, death (summed over all species)	<i>Gill</i>
b_i^p, d_i^p	<i>Per capita</i> probability of birth and death within a given timestep	<i>RF_d2S</i> , <i>RF_d2M</i> , <i>RR1S</i>
Δ_t	Size of a timestep (fixed value)	<i>RF_d2S</i> , <i>RF_d2M</i> , <i>RR1S</i>
\mathcal{E}	List of (<i>site</i> , <i>eventType</i>) pairs	<i>RR1S</i>
\mathcal{C}	List of (<i>species</i> , <i>site</i>) pairs for occupied sites	<i>RF_d2S</i> , <i>RF_d2M</i>

Algorithm B.1 *Gill* algorithm for multiple species.

```

1:  $Gill(t_{max}, A, \mathcal{S}, b^r, d^r, N)$ 
2:  $t \leftarrow 0.0$ 
3: while  $t < t_{max}$  do                                     ► Keep simulating until  $t_{max}$  reached
4:    $\lambda \leftarrow 0.0, lwb \leftarrow 0.0, p \leftarrow rand()$ 

5:   for  $i \in \mathcal{S}$  do                                         ► Set up rates
6:      $B_i^r \leftarrow b_i^r N_i$ 
7:      $D_i^r \leftarrow d_i^r N_i$ 
8:      $\lambda \leftarrow \lambda + B_i^r + D_i^r$ 
9:   end for

10:  for  $i \in \mathcal{S}$  do                                     ► Simulate next event; loop over species, exiting on event
11:     $\delta \leftarrow D_i^r / \lambda$ 
12:     $\beta \leftarrow B_i^r / \lambda$ 
13:    if  $p \leq \delta + lwb$  then
14:       $doDeath(i, A, N)$                                      ► Kill random member of species  $i$ 
15:      break                                                 ► Exit for loop
16:    end if
17:     $lwb \leftarrow lwb + \delta$ 
18:    if  $p \leq \beta + lwb$  then
19:       $doBirth(i, A, N)$                                      ► Random member of species  $i$  attempts birth
20:      break                                                 ► Exit for loop
21:    end if
22:     $lwb \leftarrow lwb + \beta$ 
23:  end for

24:   $\tau \leftarrow -\ln(rand()) / \lambda$                        ► Simulate elapsed time
25:   $t \leftarrow t + \tau$                                        ► Update time

26: end while

```

uniformly random number $p \in [0, 1)$ (generated in line 4), executing a single event if this value plus the current lower bound *lwb* is greater than or equal to p (a death at line 14 or a birth at line 19). When an event takes place the `for` loop is exited via a call to `break` (ensuring that exactly one event is executed per iteration of the `while` loop). The time to the next event τ is then computed by drawing from an exponential distribution with mean λ (line 24) and time t increased by this amount (line 25).

The call to `doDeath(i, A, N)` (line 14) randomly selects a site in A occupied by an organism of species i , sets this site to be empty and decrements the population count N_i . Similarly, the call to `doBirth(i, A, N)` (line 19) randomly selects a site in A occupied by an organism of species i and chooses with uniform probability between neighbouring sites; if the chosen site is unoccupied a birth takes place into this site and the population count N_i is incremented.

Note that if the site into which a birth is due to take place is already occupied, the birth event does not take place. The same is true of all the algorithms presented here.

This algorithm is based upon the Direct Method (DM) (Gillespie, 1977). However, because different organisms have different birth rates as a function of the occupancy of their neighbourhood, the original DM method would have required computing the birth and death event rate for every organism on the grid (i.e. $2 \times N$ rates where N is the total number of organisms). The birth rate of each organism also depends upon the occupancy of its neighbourhood, and with the Moore Neighbourhood, computing these rates entails assessing the occupancy of the 8 neighbouring cells. These would then need to be updated after an event took place (Potential ways of avoiding the need to update all event rates, since only those of organisms in the direct neighbourhood of an event would be affected, are reviewed in Slepoy et al., 2008; Sanassy et al., 2014). The main innovations in the algorithm that we employ are based on the following reasoning. Firstly, all organisms of the same species have the same death rate. This means that it is not necessary to compute or store the rates for each organism individually. Instead, we can store and compute only m species-level death rates and choose the organism to which a death applies in two steps: firstly, we select a species according to the species-level death rates, and then we choose the organism to which it applies uniform randomly within the organisms of this species. When a death occurs, we update the count of the number of organisms of that species and the overall species-level death rate. Secondly, all organisms of the same species have the same constitutional birth propensity, and their actualised birth rate depends only on the probability of a potential birth occurring into a vacant cell. Focusing on a single individual, the rate at which that individual tries to give birth (its constitutional birth propensity) multiplied by its probability of actually achieving a successful birth, is equivalent to the rate at which it gives birth successfully. Since all organisms of the same species have the same birth propensity, we can first select a species according to the species-level birth propensity (given by the product of the number of organisms of that species and the constitutional birth propensity of the species), select an organism of that

species uniform randomly, select its birth site uniform randomly from those in its neighbourhood and only implement birth should that site prove to be unoccupied. This means that we never have to compute (or store) birth rates as a function of the occupancy of the neighbourhood, and in fact requires us to keep track of only the number of organisms of each species (a vector of length m) and the site at which each resides (a vector of length N , where N is the total number of organisms in the system).

As a result, the space complexity of the algorithm scales as $O(n^2)$ in the size of the lattice (where n is the number of sites per side) and $O(m)$ in the number of species. The number of operations required per (actualised) event depends on the overall occupancy: at high occupancy, birth events are generated but not effected as births are due to take place into occupied sites. However, these probabilities are independent of the size of the lattice n , so are constant in n per actualised event. The number of operations per proposed event scales as $O(m)$ in the number of species (species selection), as $O(n^2)$ in the size of the lattice (in the worst case of a single species and full grid, we have to loop over all organisms of that species, i.e. all sites), but is constant in the number of sites in the neighbourhood as occupancy is only computed for the site into which birth is due to take place.

Algorithm B.2 RF_d2S algorithm for multiple species.

```

1:  $RF_d2S(t_{max}, A, \Delta_t, b^p, d^p)$ 
2:  $t \leftarrow 0.0$ 
3: while  $t < t_{max}$  do                                     ► Keep simulating until  $t_{max}$  reached
4:    $B \leftarrow A$ 
5:    $C \leftarrow getSites(A)$ 
6:    $shuffle(C)$ 

7:   for  $(i, site) \in C$  do                                     ► Consider sites for death
8:     if  $d_i^p \geq rand()$  then
9:        $doDeath(B, site)$ 
10:    end if
11:  end for

12:  for  $(i, site) \in C$  do                                     ► Consider sites for birth
13:    if  $b_i^p \geq rand()$  then
14:       $doBirth(i, B, B, site)$ 
15:    end if
16:  end for

17:   $t \leftarrow t + \Delta_t$                                      ► Update time
18:   $A \leftarrow B$                                              ► Update array for next generation
19: end while

```

B.1.2 RF_d2S algorithm

Algorithm B.2 shows RF_d2S . As with *Gill*, this algorithm takes arguments t_{max} and A , as well as the time step Δ_t and two vectors containing the *per capita* birth and death probability per time step for each species b^p and d^p . These vectors are calculated from b^r and d^r using the standard conversion from rates to probabilities: $b_i^p = 1 - e^{-\Delta_t b_i^r}$ (and analogously for d^p).

Time progresses in regular steps Δ_t (each constituting a *generation*) until t_{max} (*while* loop). At each generation, A is copied into B (line 4), then the set of occupied sites \mathcal{C} is computed as a list of (*species*, *site*) pairs and the order of sites is randomised using a Knuth shuffle¹ (lines 5-6). Lines 7-11 execute death events (all deaths are executed before moving on to births): for each (*species*, *site*) pair in \mathcal{C} , a random number is drawn and if this is less than or equal to the probability of death of the species at that site, a death is executed on B . Lines 12-16 execute birth events in a similar way.

The overloaded procedures $doDeath(A, site)$ and $doBirth(i, A, B, site)$ work as follows. The procedure call $doDeath(B, site)$ (line 9) eliminates the organism at the given site in B (note that B and A are the same until the first death has taken place). In procedure call $doBirth(i, B, B, site)$ (line 14), if the given site is not occupied by species i in B (this occurs if the organism originally at the site in A used to construct \mathcal{C} has died at lines 7-11) then nothing happens, ensuring that organisms that have died can no longer give birth; otherwise a neighbouring site σ is selected with uniform probability and if σ is unoccupied in B , a birth of species i takes place into site σ in B . On completing a generation B is copied into A (line 18).

B.1.3 $RR1S$ algorithm

$RR1S$ is presented in Algorithm B.3, and takes the same arguments as RF_d2S . At each Δ_t increment (generation) the set of all (*site*, *eventType*) pairs is calculated where *eventType* is either a birth or a death (line 4). The order of these pairs is randomised (line 5). In lines 6-16, (*site*, *eventType*) pairs are considered in turn. It is possible that a pair (*site*, *birth*) follows a pair (*site*, *death*) and the focal *site* is then unoccupied; this is tested for in line 7. For an occupied site (lines 7 to 16), we obtain the species i at the site (line 8). If the current event is a death, with probability d_i^p we call $doDeath$ with argument A , thus affecting this data structure directly. If the current event is a birth, we call $doBirth$ with probability b_i^p with A twice, so that births take place on A directly.

¹Although strictly speaking, shuffling the site ordering for death events is actually unnecessary since death events take place independently of one another.

Algorithm B.3 *RR1S* algorithm for multiple species.

```

1:  $RR1S(t_{max}, A, \Delta_t, b^p, d^p)$ 
2:  $t \leftarrow 0.0$ 
3: while  $t \leq t_{max}$  do                                     ► Keep simulating until  $t_{max}$  reached
4:    $\mathcal{E} \leftarrow getEvents(A)$                              ► Make list of  $(site, eventType)$  pairs
5:    $shuffle(\mathcal{E})$ 

6:   for  $(site, eventType) \in \mathcal{E}$  do
7:     if  $isOccupied(site, A)$  then                               ► Check if site occupied
8:        $i \leftarrow getSpecies(site, A)$                          ► If occupied, get species at site
9:       if  $eventType = death \wedge d_i^p \geq rand()$  then
10:         $doDeath(A, site)$ 
11:       end if
12:       if  $eventType = birth \wedge b_i^p \geq rand()$  then
13:         $doBirth(i, A, A, site)$ 
14:       end if
15:     end if
16:   end for

17:    $t \leftarrow t + \Delta_t$                                      ► Update time for next generation
18: end while

```

B.1.4 Multiple births *RF_d2M* and *RR1M*

The multiple births algorithms *RF_d2M* and *RR1M* differ only slightly from the single births version, and therefore only *RF_d2M* is shown in Algorithm B.4. Firstly, instead of calculating the probability of birth within a timestep, a number of births m is drawn from a Poisson distribution with mean $\Delta_t b_i^r$ (line 13) and correspondingly the algorithms take b^r as an argument (in place of b^p , line 1). Secondly, $doBirth(i, B, B, site)$ is called m times (where m represents the number of births to attempt and may be zero) at line 14-16 (if $m = 0$, the loop is never executed).

Note that Algorithm B.4 differs from Algorithm B.2 only in the way births are performed, i.e. line 14 in Algorithm B.2 performs zero or one birth and lines 14-16 in Algorithm B.4 perform zero, one or more births.

Algorithm B.4 *RF_d2M* algorithm for multiple species.

```

1: RFd2M( $t_{max}, A, \Delta_t, b^r, d^p$ )
2:  $t \leftarrow 0.0$ 
3: while  $t < t_{max}$  do                                     ► Keep simulating until  $t_{max}$  reached
4:    $B \leftarrow A$ 
5:    $\mathcal{C} \leftarrow getSites(A)$ 
6:   shuffle( $\mathcal{C}$ )

7:   for  $(i, site) \in \mathcal{C}$  do                                   ► Consider sites for death
8:     if  $d_i^p \geq rand()$  then
9:       doDeath( $B, site$ )
10:    end if
11:  end for

12:  for  $(i, site) \in \mathcal{C}$  do                                   ► Consider sites for death
13:     $m \leftarrow PoissRand(\Delta_t b_i^r)$                        ► Poisson random number of births
14:    for  $j \in [1..m]$  do
15:      doBirth( $i, B, B, site$ )   ► Birth of  $i$  if still present on  $B$  and site  $\sigma$  empty
16:    end for
17:  end for

18:   $t \leftarrow t + \Delta_t$                                      ► Update time for next generation
19:   $A \leftarrow B$ 
20: end while

```

Appendix C

Chapter 4

C.1 Alternative predictors of ∇T_m

C.1.1 p_i^π

Figure C.1 shows a quadratic relationship, explaining why the square of p_i^π gives a good prediction of ∇T_m .

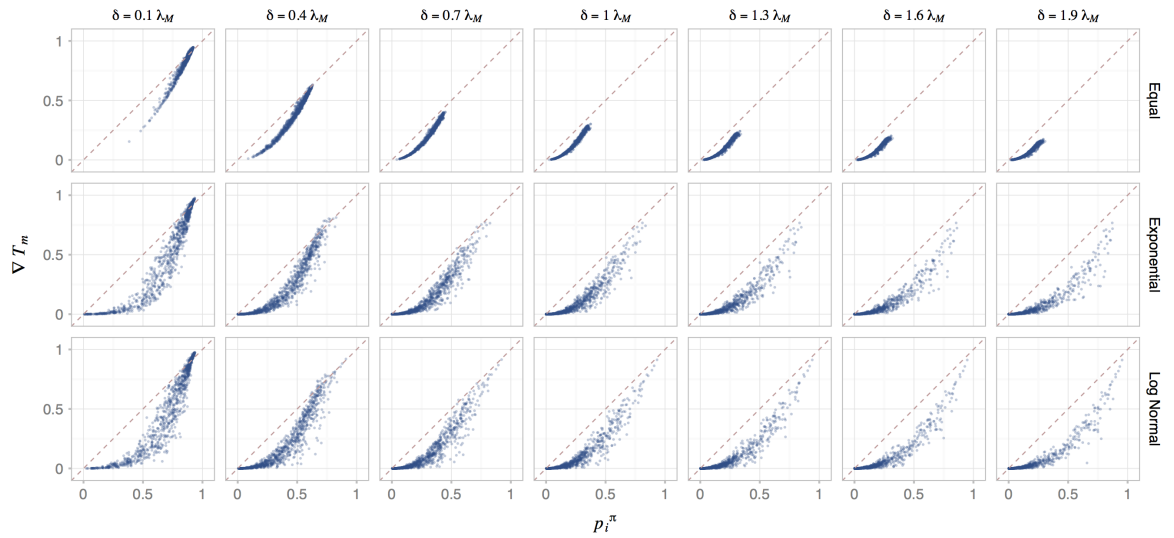


Figure C.1: The relationship between p_i^π and ∇T_m (parameters as in Figure 4.3).

C.1.2 p_i^*

In Figure C.2, we see a similar relationship between p_i^* (the probability occupancy of the patch at deterministic steady state) and ∇T_m as for in Fig. C.1 for the highest levels of

persistence; however, the p_i^* values are less predictive just above the deterministic persistence threshold (i.e. $(p_i^*)^2$ underestimates patch values). Unsurprisingly, p_i^* provides no information at all about the relative value of a patch at and below this threshold.

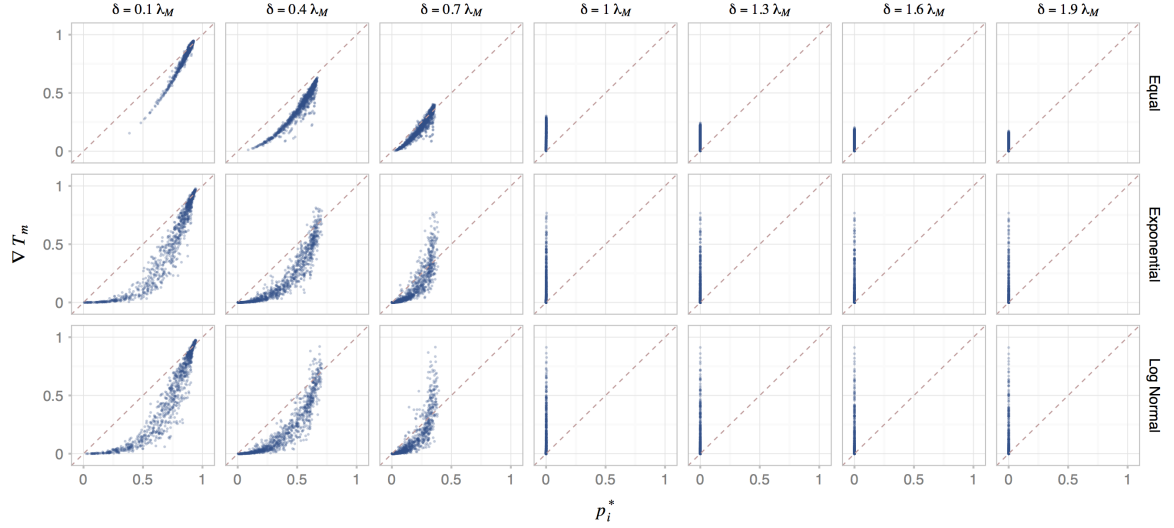


Figure C.2: The relationship between p_i^* and ∇T_m (parameters as in Figure 4.3).

C.1.3 A_i

Figure C.3 shows the predictive value of the relative area of a patch for predicting ∇T_m . For obvious reasons, in the case of equal patch areas, this measure provides no useful information about patch removal. The area of the patch shows more potential as a useful predictor below the deterministic threshold for exponentially and log normally distributed patch areas than above. This makes sense because below the deterministic persistence threshold, patch extinction events are more common than colonisation and whilst patch colonisation is dependent on the areas of other patches, extinction events depend only on the area of the patch itself. Nonetheless, as a predictor of patch value, A_i is not robust to patch area distributions. One would also expect it to make better predictions in the case where persistence is primarily due to low extinction rates, as opposed to high colonisation rates (i.e. for systems with slow patch turnover), which could be adjusted using parameter e and c as shown here, or by varying the mean migration distance $1/\alpha$.

C.1.4 V_i

Figure C.4 shows the relationship between patch values \tilde{V}_i (Hanski and Ovaskainen, 2000) and relative change in mean time to extinction. The figures show that for equal patch areas,

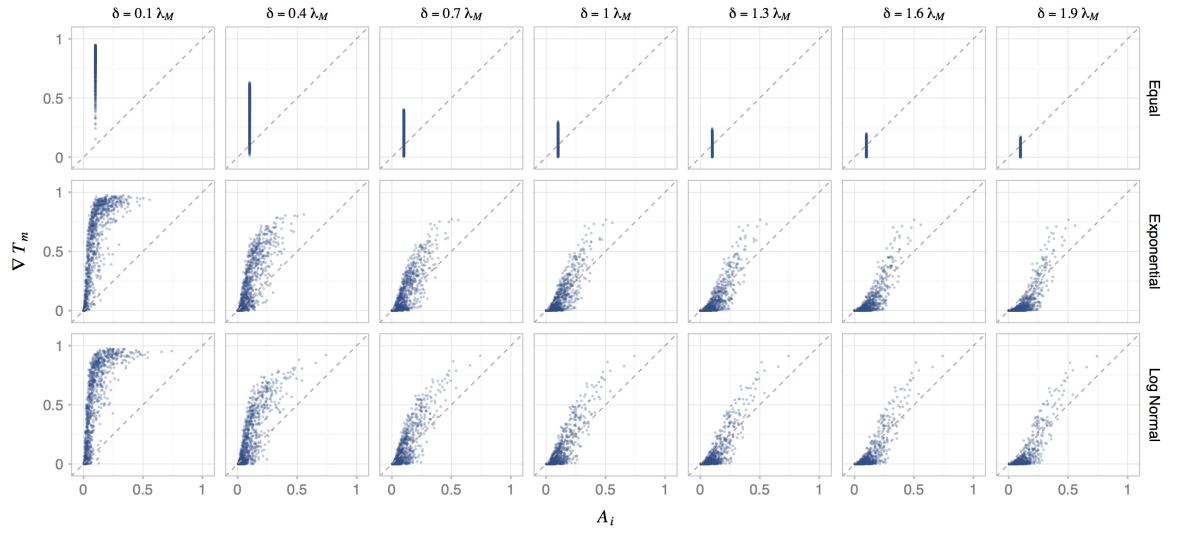


Figure C.3: The relationship between A_i and ∇T_m (parameters as in Figure 4.3).

patch values tend to cluster into two groups of those of about $\tilde{V}_i \approx 0$ and those of about $\tilde{V}_i \approx 0.5$. For exponentially and log normally distributed patch areas, \tilde{V}_i values are demonstrate greater spread and there is a positive correlation between patch values and the effect of their removal on metapopulation persistence. However, except at the two highest levels of persistence, \tilde{V}_i underestimates the effect of removal on T_m , and there is considerable scatter for all persistence levels.

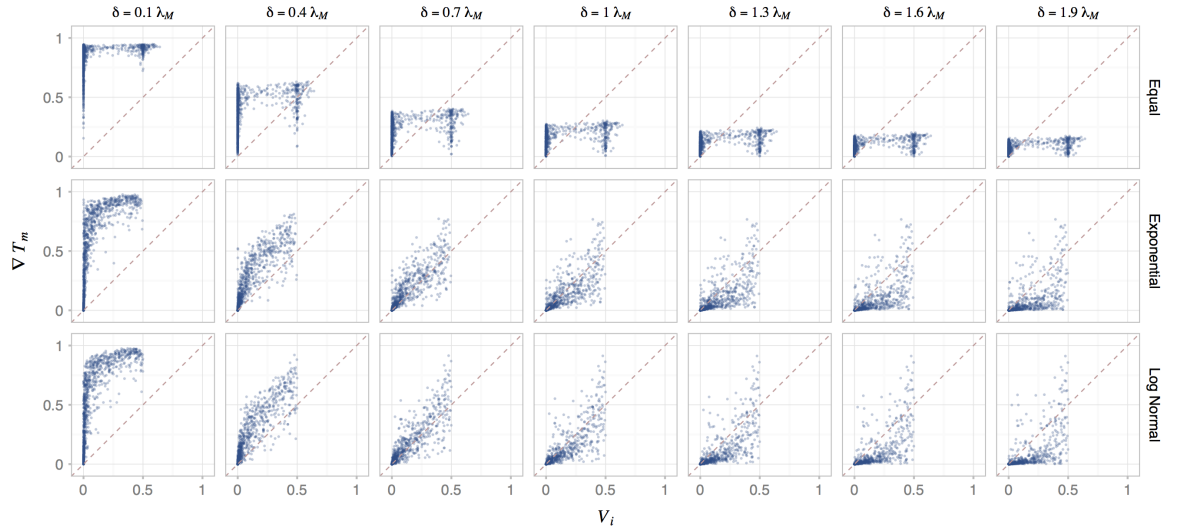


Figure C.4: The relationship between V_i and ∇T_m (parameters as in Figure 4.3).

Note that the values plotted here are approximate \tilde{V}_i values from the matrix \mathbf{M} rather than the ‘true’ V_i values (the distinction is explained in the main text).

C.2 Alternative predictors of ∇S^π

C.2.1 p_i^π and proportion of S^π due to patch i

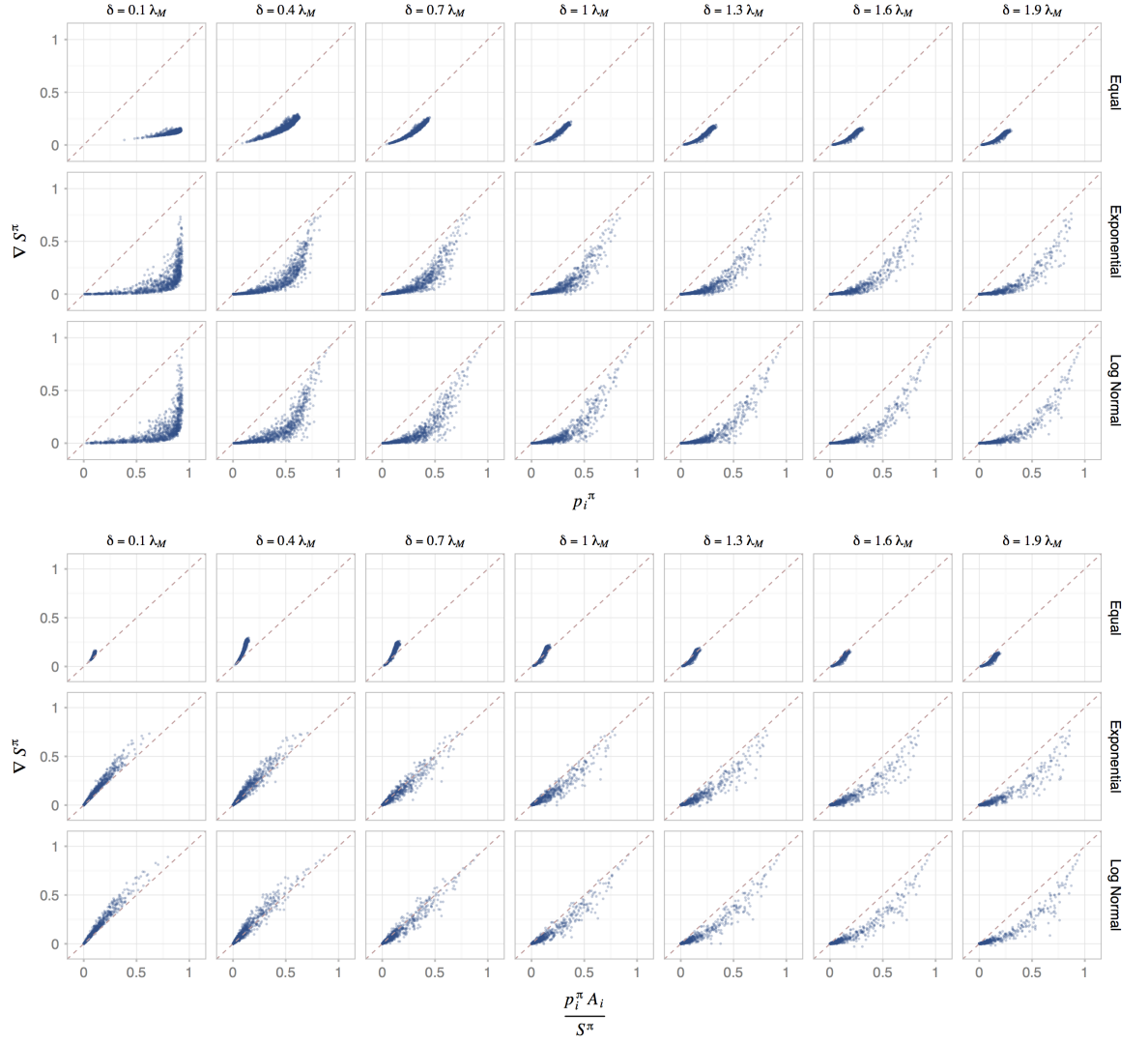
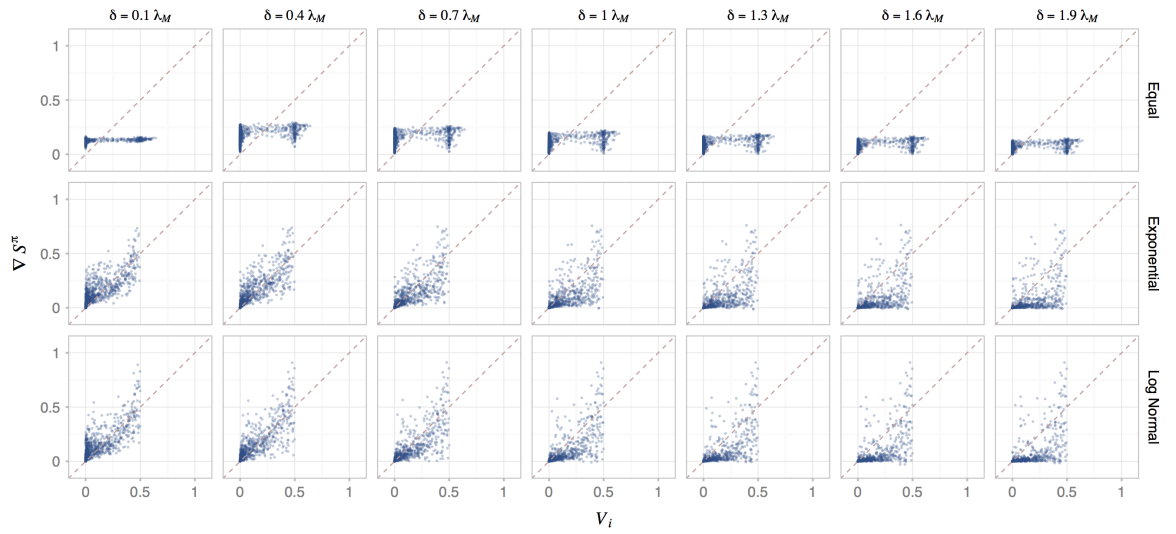
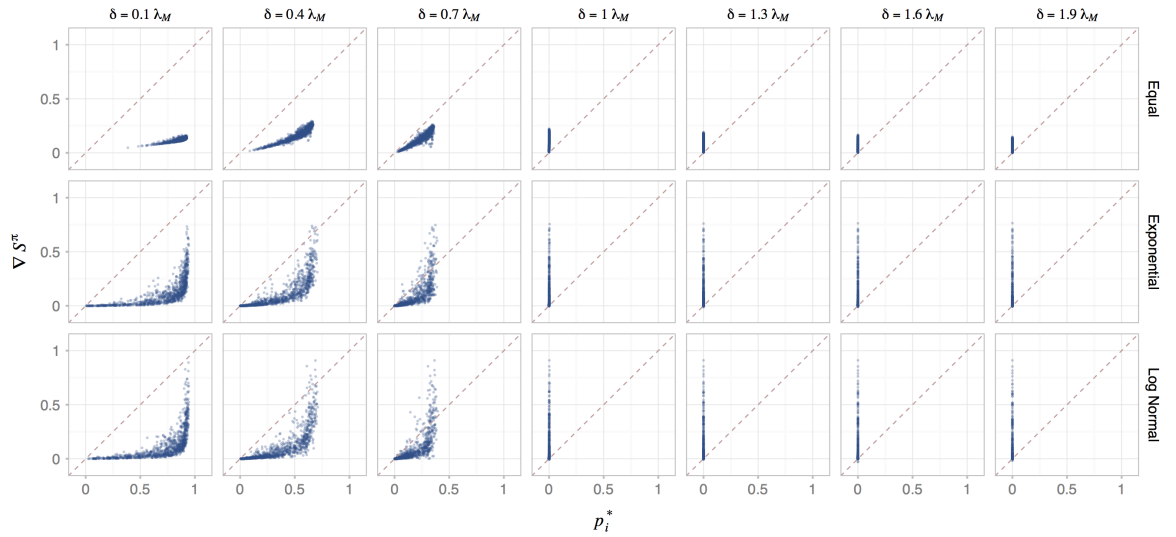


Figure C.5: The relationship between potential measures using the raw p_i^π values and S^π .

In Figure C.5 we see a similar quadratic relationship between these predictors and S^π as previously with T_m .

C.2.2 V_i and V_i^π

As was the case for T_m , the predictor \tilde{V}_i makes a poor job of approximating the S^π values.

Figure C.6: The relationship between \tilde{V}_i and S^π .C.2.3 p^* Figure C.7: The relationship between p^* and S^π .

As expected, the predictor p^* provides no information below the persistence threshold. Above this threshold, it has a similar shape to p^π but is overall less accurate.

Appendix D

Chapter 5

D.1 QSD comparison statistics

We tested four statistics for comparing the exact QSD with the approximations. The QSD of the SRLM can be thought of as a vector of probabilities of length 2^n equal to the number of states, where n is the number of patches. In order to test whether the different approximations improved or became worse as n increased, we required a statistic that measured the distance between two approximations in a way that was comparable between systems with different numbers of patches. A number of measures have been suggested for comparing probability distributions (Cha, 2007).

In order to establish the most appropriate statistic for our purposes, we tested the following four statistics: the Pearson r correlation coefficient, the root mean squared error (rmse), the angle between the two vectors described by the two approximations, and the Kullback-Leibner divergence, each described below. To achieve this, for each n , we first simulated 100 vectors of length 2^n by drawing a value in the interval $[0, 1)$ for each entry. We then normalised the vectors by dividing by the sum of the elements to give a uniform random probability vector. Every vector in the set was compared to every other vector using each of the statistics selected. Findings are presented below. We selected the Pearson correlation coefficient as this measure did not suffer from the limitation of bias and thus allowed comparisons between QSD accuracy at different values of n .

In the definitions provided below, we denote the true distribution by X and the approximation by Y .

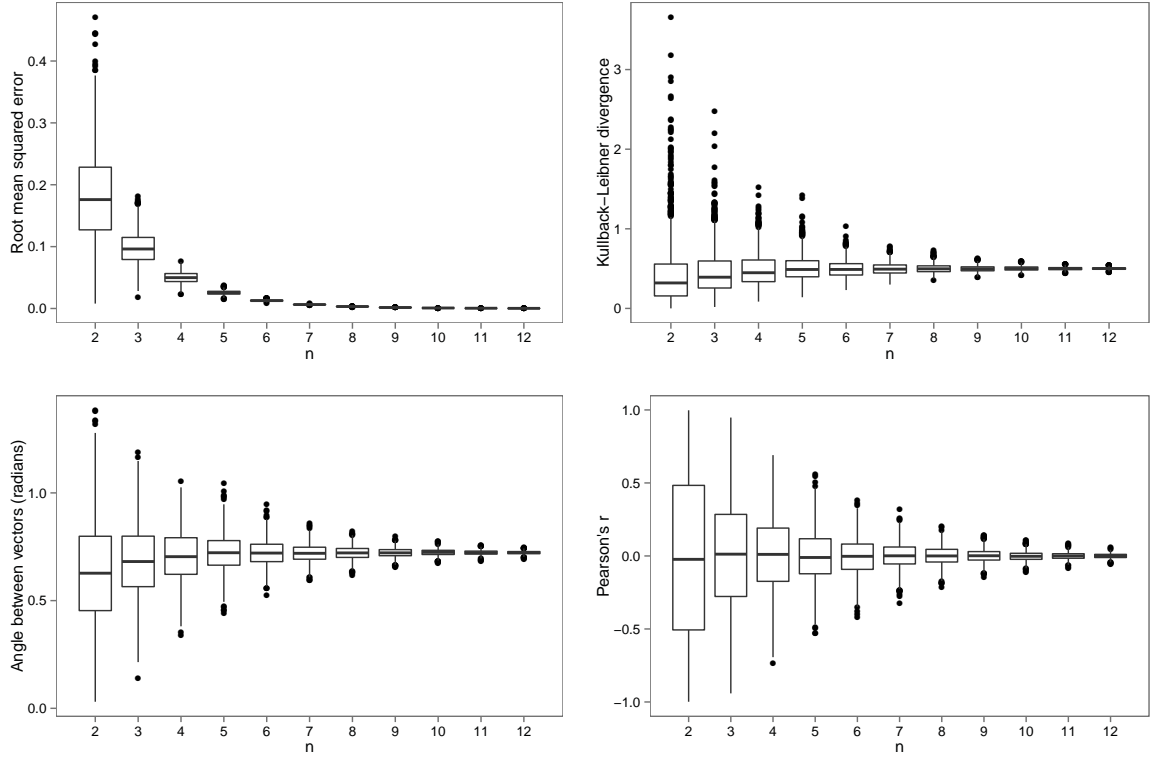


Figure D.1: Box plots showing the statistics considered for vectors of length 2^n .

D.1.1 Pearson correlation coefficient

The Pearson product moment correlation coefficient, usually denoted r , measures the strength of the linear relationship between two variables. Values of the coefficient range from -1 , when there is a perfect negative linear relationship, to $+1$ when there is a perfect positive linear relationship; a value of 0 indicates that there is no linear relationship between the two variables. It is defined as:

$$r = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\sum_{i=1}^l (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^l (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^l (Y_i - \bar{Y})^2}}.$$

In the case of perfect correspondence between the two representations of the QSD, we would obtain $r = 1$ (in fact, the state probabilities would in fact lie on the line $y = x$ although Pearson's correlation coefficient does not test the slope of the line).

D.1.2 Root mean square error

The root mean square error (rmse) or root mean square deviation (rmsd) is often used to compare values predicted by a model and those that are actually observed. In the case of

perfect correspondence, its value is zero. It is defined as:

$$rmse = \sqrt{\frac{\sum_{i=1}^L (X_i - Y_i)^2}{L}}.$$

It measures the mean of the square of the distance between the predicted value and the observed value, and is rescaled by taking the square root to measure the mean distance in the original units. Although the rmse is often used for the purpose of comparing observed values to those predicted by a model, it has the disadvantage of being scale dependent. (Note that the rmse is robust to changes in the number of states when used to compare non-normalised vectors. Here we are normalising to generate a probability distribution; doing so reduces the degrees of freedom of the distribution.)

D.1.3 Angle between vectors

If the vectors representing two distributions are the same, then the angle between those vectors (in 2^n -dimensional space) should be zero. The angle between the two vectors can be found from the dot product or scalar product. The dot product of two vectors is defined as

$$X \cdot Y = ||X|| ||Y|| \cos(\theta)$$

in which $||X||$ represents the magnitude of the vector X and θ the angle between the vectors. Rearranging the formula, and calculating the magnitude of the vector in the usual way, gives the angle. This measure corresponds to Equation 26 in Cha (2007).

D.1.4 Kullback-Leibler divergence

The Kullback-Leibler divergence or *discrimination information* (Kullback and Leibler, 1951; Kullback, 1987) is a non-parametric measure of the difference between two probability density distributions, P and Q , used in information theory and probability theory. For discrete probability distributions, it is defined as

$$D_{KL}(P||Q) = \mathbb{E} \left[\log \left(\frac{p(x)}{q(x)} \right) \right].$$

More specifically, it measures the divergence of Q from P as the information that is lost when Q is used to approximate P . Put another way, this corresponds to the expected number of additional bits required to code samples from P when using a code based on Q , rather than using a code based on P directly. Typically, one thinks of P as the ‘true’ distribution of data, observations, or a precisely calculated theoretical distribution whereas Q denotes a

theory, model, description, or approximation of P . This measure corresponds to Equation 48 in Cha (2007).

D.1.5 Findings

Of the statistics tested, only Pearson’s correlation coefficient was shown to be unbiased in the sense of giving the same mean value independent of the number of patches. However, the variance was higher for smaller n .

D.2 Robustness testing

We include here additional figures representing robustness testing, referred to in the main text.

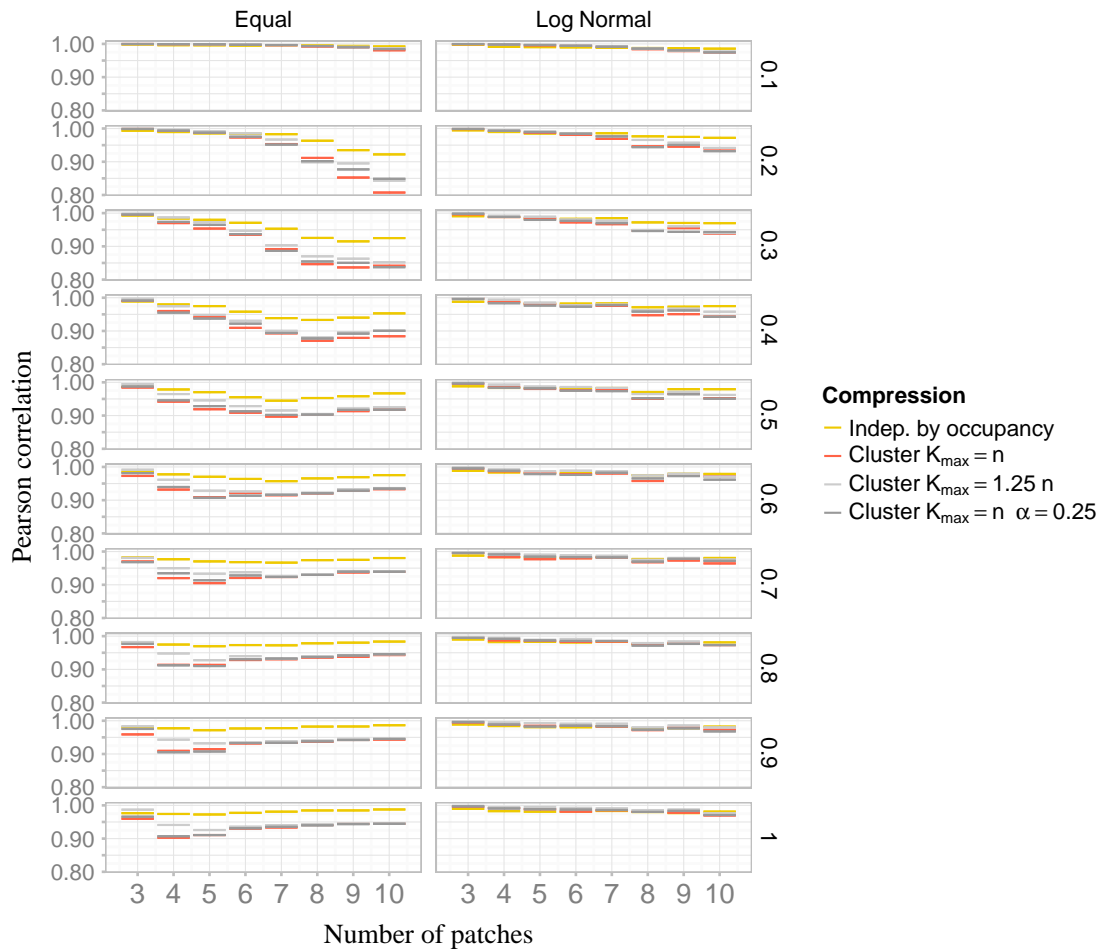


Figure D.2: Comparison of adjustments to clustering parameters with $K_{max} \approx n$.

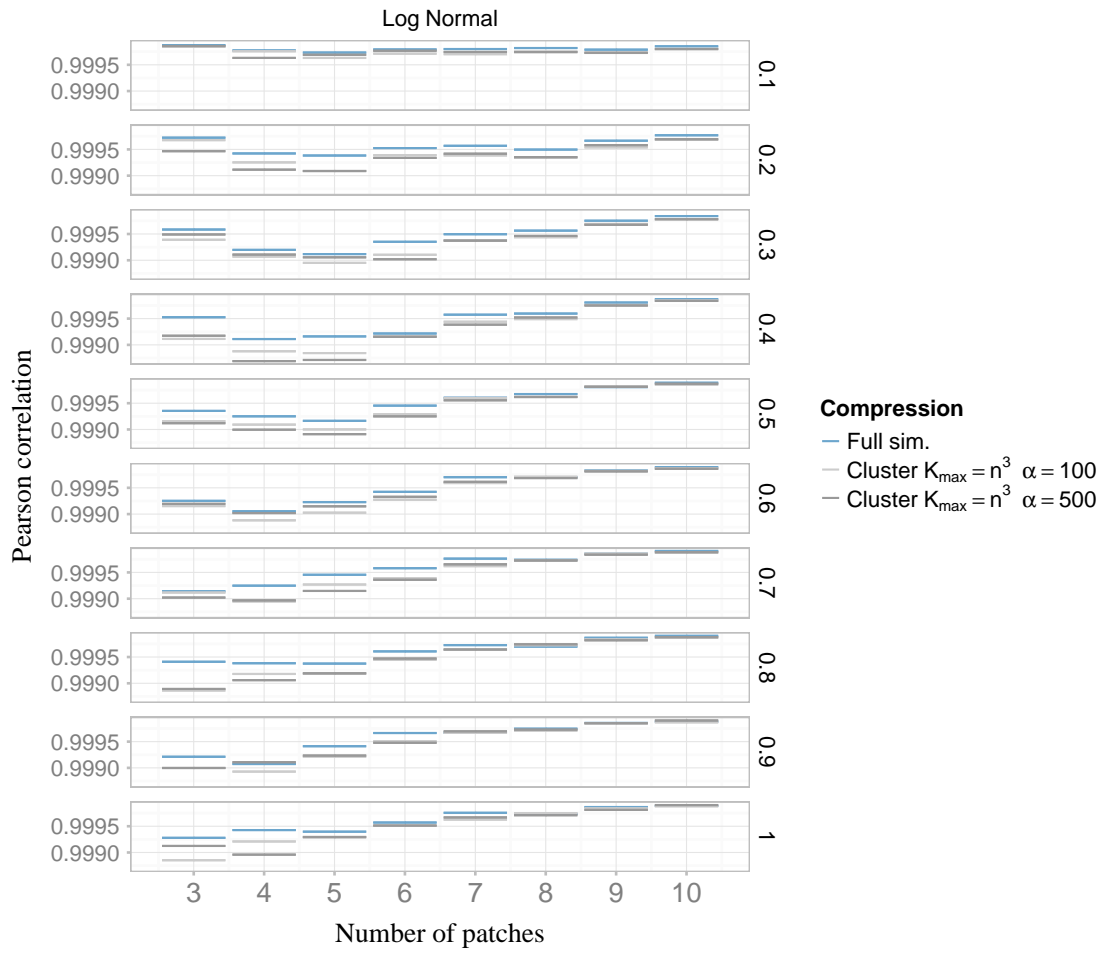


Figure D.3: Algorithm accuracy for clustering algorithm with n^3 and different values of α .

Appendix E

Chapter 6

E.1 Correspondence between Levins and logistic growth model

The following manipulation shows that the re-parameterised Levins model is equivalent to the original model. The first line is found by substituting $K = 1 - r/c$, $\rho = c - r$ and $p = N/K$ into the standard logistic model $\frac{dN}{dt} = \rho N \left(1 - \frac{N}{K}\right)$.

$$dp/dt = (c - r)p \left[1 - \frac{p}{1 - \frac{r}{c}}\right] \quad (\text{E.1})$$

$$= (c - r)p \left[1 - \frac{cp}{c - r}\right] \quad (\text{E.2})$$

$$= \left[(c - r) - \frac{(c - r)cp}{c - r}\right] p \quad (\text{E.3})$$

$$= [(c - r) - cp] p \quad (\text{E.4})$$

$$= [c - r - cp] p \quad (\text{E.5})$$

$$= [c(1 - p) - r] p \quad (\text{E.6})$$

$$= cp(1 - p) - rp. \quad (\text{E.7})$$

The last line is the standard formulation of the Levins model from Equation 6.2.

E.2 Derivation of non-standard form of *SIIS* Levins model

We proceed by analogy with the derivation in E.1.

$$\frac{dp_1}{dt} = c_1 \left(1 - \left[1 + \frac{r_1}{r_2} \right] p_1 \right) p_1 + c_2 \left(1 - \left[1 + \frac{r_1}{r_2} \right] p_1 \right) \frac{r_1}{r_2} p_1 - r_1 p_1 \quad (\text{E.8})$$

$$= \left(c_1 + \frac{r_1}{r_2} c_2 \right) \left(1 - \left[1 + \frac{r_1}{r_2} \right] p_1 \right) p_1 - r_1 p_1 \quad (\text{E.9})$$

$$= \left[\left(c_1 + c_2 \frac{r_1}{r_2} \right) \left(1 - \left(1 + \frac{r_1}{r_2} \right) p_1 \right) - r_1 \right] p_1 \quad (\text{E.10})$$

$$= \left[\left(c_1 + c_2 \frac{r_1}{r_2} \right) - r_1 - \left(c_1 + c_2 \frac{r_1}{r_2} \right) \left(1 + \frac{r_1}{r_2} \right) p_1 \right] p_1 \quad (\text{E.11})$$

$$= \left[\left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right) - \left(c_1 + c_2 \frac{r_1}{r_2} \right) \left(1 + \frac{r_1}{r_2} \right) p_1 \right] p_1. \quad (\text{E.12})$$

We are now searching for a factor that we can pull out of the expression that represents the rate of colonisation minus the rate of recovery. We proceed as follows by multiplying the top and bottom of the right-hand term in the bracket by $\left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right)$:

$$= \left[\left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right) - \frac{\left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right) \left(c_1 + c_2 \frac{r_1}{r_2} \right) \left(1 + \frac{r_1}{r_2} \right) p_1}{\left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right)} \right] p_1. \quad (\text{E.13})$$

$$= \left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right) p_1 \left[1 - \frac{\left(c_1 + c_2 \frac{r_1}{r_2} \right) \left(1 + \frac{r_1}{r_2} \right) p_1}{\left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right)} \right] \quad (\text{E.14})$$

$$= \left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right) p_1 \left[1 - \frac{\left(1 + \frac{r_1}{r_2} \right) p_1}{\frac{\left(c_1 + c_2 \frac{r_1}{r_2} - r_1 \right)}{\left(c_1 + c_2 \frac{r_1}{r_2} \right)}} \right] \quad (\text{E.15})$$

$$= \left(c_1 + \frac{r_1}{r_2} c_2 - r_1 \right) p_1 \left[1 - \frac{\left(1 + \frac{r_1}{r_2} \right) p_1}{1 - \frac{r_1 r_2}{c_1 r_2 + r_1 c_2}} \right]. \quad (\text{E.16})$$

The transition between the penultimate and last line is simply based upon a re-arrangement of the denominator of the fraction.

Bibliography

- Adler, F. and B. Nuernberger, 1994. Persistence in Patchy Irregular Landscapes. *Theoretical Population Biology* **45**:41–75.
- Alden, K., J. Timmis, P. S. Andrews, H. Veiga-Fernandes, and M. C. Coles, 2012. Pairing experimentation and computational modeling to understand the role of tissue inducer cells in the development of lymphoid organs. *Frontiers in Immunology* **3**:1–20.
- Aldous, D., B. Flannery, and J. L. Palacios, 1988. Two Applications of Urn Processes The Fringe Analysis of Search Trees and The Simulation of Quasi-Stationary Distributions of Markov Chains. *Probability in the Engineering and Informational Sciences* **2**:293.
- Allee, W. C., 1931. Animal Aggregations. A study in General Sociology. The University of Chicago Press, Chicago, IL.
- Alonso, D. and A. McKane, 2002. Extinction dynamics in mainland-island metapopulations: an N-patch stochastic model. *Bulletin of mathematical biology* **64**:913–58.
- Amarasekare, P., 1998. Allee effects in metapopulation dynamics. *The American Naturalist* **152**:298–302.
- Andrews, P. S., S. Stepney, T. Hoverd, F. A. C. Polack, A. T. Sampson, and J. Timmis, 2011. CoSMoS process, models, and metamodels. In S. Stepney, P. H. Welch, P. S. Andrews, and C. G. Ritson, editors, *CoSMoS 2011*, pages 1–14, Frome, UK. Department of Computer Science, University of York, Luniver Press.
- Artalejo, J. R., 2012. On the time to extinction from quasi-stationarity: A unified approach. *Physica A: Statistical Mechanics and its Applications* **391**:4483–4486.
- Artalejo, J. R. and M. J. Lopez-Herrero, 2010. Quasi-stationary and ratio of expectations distributions: a comparative study. *Journal of theoretical biology* **266**:264–74.
- Ashford, R. W., 1997. What it takes to be a reservoir host. *Belgian Journal of Zoology* **127**:85–90.

- Ashford, R. W., 2003. When Is a Reservoir Not a Reservoir? *Emerging infectious diseases* **9**:1495–1496.
- Balcan, D., V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America* **106**:21484–9.
- Barbour, A. D., R. McVinish, and P. K. Pollett, 2014. Connecting deterministic and stochastic metapopulation models. Technical report, Universitat Zurich and University of Queensland.
- Barrat, A., M. Barthlemy, and A. Vespignani, 2008. Dynamical Processes on Complex Networks. Cambridge University Press, first edition.
- Bartlett, M. S., 1956. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 4, pages 81–109. University of California Press Berkeley.
- Bartlett, M. S., 1957a. Measles Periodicity and Community Size. *Journal of the Royal Statistical Society* **120**:48–70.
- Bartlett, M. S., 1957b. On theoretical models for competitive and predatory biological systems. *Biometrika* **44**:27–42.
- Bartlett, M. S., 1960. Stochastic population models in ecology and epidemiology. Methuen, London, United Kingdom.
- Berec, L., 2002. Techniques of spatially explicit individual-based models: construction, simulation, and mean-field analysis. *Ecological Modelling* **150**:55–81.
- Berec, L., E. Angulo, and F. Courchamp, 2007. Multiple Allee effects and population management. *Trends in Ecology & Evolution* **22**:185–91.
- Best, A., S. Webb, A. White, and M. Boots, 2011. Host resistance and coevolution in spatially structured populations. *Proceedings of the Royal Society B: Biological Sciences* **278**:2216–22.
- Beyer, H. L., K. Hampson, T. Lembo, S. Cleaveland, M. Kaare, and D. T. Haydon, 2012. The implications of metapopulation dynamics on the design of vaccination campaigns. *Vaccine* **30**:1014–22.
- Bhatt, S., P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay, 2013. The global distribution and burden of dengue. *Nature* **496**:504–507.

- Billings, L., L. Mier-y Teran-Romero, B. Lindley, and I. B. Schwartz, 2013. Intervention-based stochastic disease eradication. *PLoS One* **8**:e70211.
- Birch, C. P., 2006. Diagonal and orthogonal neighbours in grid-based simulations: Buffon's stick after 200 years. *Ecological Modelling* **192**:637–644.
- Birch, C. P., S. P. Oom, and J. A. Beecham, 2007. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling* **206**:347–359.
- Black, A. J. and A. J. McKane, 2012. Stochastic formulation of ecological models and their applications. *Trends in Ecology & Evolution* **27**:337–345.
- Blanchet, J., P. Glynn, and S. Zheng, 2014. Theoretical analysis of a stochastic approximation approach for computing quasi-stationary distributions. Technical report, Columbia University, New York, USA.
- Bolker, B., 1997. Using Moment Equations to Understand Stochastically Driven Spatial Pattern Formation in Ecological Systems. *Theoretical Population Biology* **52**:179–197.
- Bolker, B. M., 2008. Ecological Models and Data in R. Princeton University Press.
- Boukal, D. S. and L. Berec, 2002. Single-species Models of the Allee Effect : Extinction Boundaries, Sex Ratios and Mate Encounters. *Journal of Theoretical Biology* **218**:375–394.
- Brassil, C. E., 2001. Mean time to extinction of a metapopulation with an Allee effect. *Ecological Modelling* **143**:9–16.
- Breckling, B., G. Pe'er, and Y. G. Matsinos, 2011. Cellular Automata in Ecological Modelling. In F. Jopp, H. Reuter, and B. Breckling, editors, *Modelling Complex Ecological Dynamics*, pages 105–117. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Cabeza, M. and A. Moilanen, 2001. Design of reserve networks and the persistence of biodiversity. *Trends in Ecology & Evolution* **16**:242–248.
- Cao, Y., D. T. Gillespie, and L. R. Petzold, 2006. Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics* **124**:044109.
- Caron-Lormier, G., R. W. Humphry, D. A. Bohan, C. Hawes, and P. Thorbek, 2008. Asynchronous and synchronous updating in individual-based models. *Ecological Modelling* **212**:522–527.
- Case, T. J., 1999. An Illustrated Guide to Theoretical Ecology. Oxford University Press, USA, New York, USA.

- Castle, M. D. and C. A. Gilligan, 2012. An Epidemiological Framework for Modelling Fungicide Dynamics and Control. *PLoS One* **8**:1–10.
- Caswell, H., 1988. Theory and models in ecology: a different perspective. *Ecological Modelling* **43**:33–44.
- Caughley, G., 1994. Directions in Conservation Biology. *The Journal of Animal Ecology* **63**:215.
- Cha, S.-H., 2007. Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences* **1**:300–307.
- Cobbold, C. A. and F. Lutscher, 2014. Mean occupancy time: linking mechanistic movement models, population dynamics and landscape ecology to population persistence. *Journal of mathematical biology* **68**:549–79.
- Cohen, M. L., 2000. Changing patterns of infectious disease. *Nature* **406**:762–7.
- Collet, P., S. Martínez, S. Méléard, and J. San Martín, 2010. Quasi-stationary distributions for structured birth and death processes with mutations. *Probability Theory and Related Fields* **151**:191–231.
- Conlan, A. J. K., P. Rohani, A. L. Lloyd, M. Keeling, and B. T. Grenfell, 2010. Resolving the impact of waiting time distributions on the persistence of measles. *Journal of the Royal Society, Interface / the Royal Society* **7**:623–40.
- Cornforth, D., D. G. Green, D. Newth, and M. Kirley, 2002. Do Artificial Ants March in Step? Ordered Asynchronous Processes and Modularity in Biological Systems. In R. Standish, H. Abbass, and M. Bedau, editors, *Artificial Life VIII*, number 1988, pages 28–32. MIT Press.
- Cox, D. R. and H. D. Miller, 1965. The Theory of Stochastic Processes. Chapman and Hall, London, United Kingdom.
- Cunniffe, N. J. and C. A. Gilligan, 2010. Invasion , persistence and control in epidemic models for plant pathogens : the effect of host demography. *Journal of The Royal Society Interface* **7**:439–451.
- Daily, G. C., S. Polasky, J. Goldstein, P. M. Kareiva, H. a. Mooney, L. Pejchar, T. H. Ricketts, J. Salzman, and R. Shallenberger, 2009. Ecosystem services in decision making: time to deliver. *Frontiers in Ecology and the Environment* **7**:21–28.

- Darroch, J. N. and E. Seneta, 1967. On Quasi-Stationary Distributions in Absorbing Continuous-Time Finite Markov Chains. *Journal of Applied Probability* **4**:192–196.
- Daszak, P., A. A. Cunningham, and A. D. Hyatt, 2001. Anthropogenic environmental change and the emergence of infectious diseases in wildlife. *Acta Tropica* **78**:103–16.
- Day, J. and H. Possingham, 1995. A Stochastic Metapopulation Model with Variability in Patch Size and Position. *Theoretical Population Biology* **48**:333–360.
- de Oliveira, M. and R. Dickman, 2005. How to simulate the quasistationary state. *Physical Review E* **71**:016129.
- De Oliveira, M. M. and R. Dickman, 2006. Quasi-Stationary Simulation : the Subcritical Contact Process. *Brazilian Journal of Physics* **36**:685–689.
- Denning, P., 2007. An interview with peter denning on the great principles of computing. *Ubiquity* pages 1–9.
- Denning, P. J., 2003. The great principles of computing. *Communications of the ACM* **98**:369–372.
- Denning, P. J., 2005. Is computer science science? *Communications of the ACM* **48**:27.
- Denning, P. J., 2013. The science in computer science. *Communications of the ACM* **56**:35.
- Denning, P. J., D. E. Comer, D. Gries, M. C. Mulder, A. Tucker, a. J. Turner, and P. R. Young, 1989. Computing as a Discipline. *Communications of the ACM (CACM)* **32**:9–23.
- Denning, P. J. and P. A. Freeman, 2009. The profession of IT: computing’s paradigm. *Communications of the ACM* **52**:28–31.
- Denning, P. J. and P. S. Rosenbloom, 2009. Computing: the fourth great domain of science. *Communications of the ACM* **52**:27.
- Diekmann, O., J. Heesterbeek, and J. Metz, 1990. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology* **28**:365–382.
- Drake, J. M., J. Shapiro, and B. D. Griffen, 2011. Experimental demonstration of a two-phase population extinction hazard. *Journal of the Royal Society, Interface / the Royal Society* **8**:1472–9.
- Durrett, R. and S. Levin, 1994. The Importance of Being Discrete (and Spatial). *Theoretical Population Biology* **46**:363–394.

- Feng, Z. and Y. DeWoody, 2004. Conservation thresholds derived from metapopulation models. In R. K. Swihart and J. E. Moore, editors, *Conserving Biodiversity in Agricultural Landscapes: Model-Based Planning Tools*, pages 49–68. Purdue University Press.
- Fischer, M., M. Van Kleunen, and B. Schmid, 2008. Genetic Allee effects on performance, plasticity and developmental stability in a clonal plant. *Ecology Letters* **3**:530–539.
- Fleurence, R. L. and C. S. Hollenbeak, 2007. Rates and probabilities in economic modelling: transformation, translation and appropriate application. *Pharmacoeconomics* **25**:3–6.
- Fox, R. E., X. Zhong, S. M. Krone, and E. M. Top, 2008. Spatial structure and nutrients promote invasion of IncP-1 plasmids in bacterial populations. *The ISME journal* **2**:1024–39.
- Frank, K., 2005. Metapopulation persistence in heterogeneous landscapes: lessons about the effect of stochasticity. *The American Naturalist* **165**:374–88.
- Freeman, P. A., 2008. Back to experimentation. *Communications of the ACM* **51**:21.
- Gertsev, V. and V. Gertseva, 2004. Classification of mathematical models in ecology. *Ecological Modelling* **178**:329–334.
- Gillespie, D. T., 1977. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81**:2340–2361.
- Gillespie, D. T., 2001. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics* **115**:1716.
- Gilligan, C. A. and F. van den Bosch, 2008. Epidemiological models for invasion and persistence of pathogens. *Annual Review of Phytopathology* **46**:385–418.
- Gonzalez, A., 2013. The ecological deficit. *Nature* **503**:206–7.
- Groisman, P. and M. Jonckheere, 2012. Simulation of quasi-stationary distributions on countable spaces. *arxiv* pages 1–22.
- Gubbins, S., C. A. Gilligan, and A. Kleczkowski, 2000. Population dynamics of plant-parasite interactions: thresholds for invasion. *Theoretical population biology* **57**:219–233.
- Gyllenberg, M. and D. S. Silvestrov, 1994. Quasi-stationary distributions of a stochastic metapopulation model. *Journal of Mathematical Biology* **33**:35–70.
- Hagenaars, T. J., C. A. Donnelly, and N. M. Ferguson, 2004. Spatial heterogeneity and the persistence of infectious diseases. *Journal of Theoretical Biology* **229**:349–59.

- Hanski, I., 1982. Dynamics of regional distribution: the core and satellite species hypothesis. *Oikos* **38**:210–221.
- Hanski, I. and M. Gyllenberg, 1997. Uniting Two General Patterns in the Distribution of Species. *Science (New York, N.Y.)* **275**:397–400.
- Hanski, I. and O. Ovaskainen, 2000. The metapopulation capacity of a fragmented landscape. *Nature* **404**:755–758.
- Harnik, P. G., C. Simpson, and J. L. Payne, 2012. Long-term differences in extinction risk among the seven forms of rarity. *Proceedings of the Royal Society B: Biological Sciences* **279**:4969–76.
- Hastings, A., 2004. Transients: the key to long-term ecological understanding? *Trends in Ecology & Evolution* **19**:39–45.
- Hastings, J., K. Haug, and C. Steinbeck, 2014. Ten recommendations for software engineering in research. *GigaScience* pages 1–4.
- Hay, S. I., C. A. Guerra, P. W. Gething, A. P. Patil, A. J. Tatem, A. M. Noor, C. W. Kabaria, B. H. Manh, I. R. F. Elyazar, S. Brooker, D. L. Smith, R. A. Moyeed, and R. W. Snow, 2009. A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Medicine* **6**:e1000048.
- Haydon, D. T., S. Cleaveland, L. H. Taylor, and M. K. Laurenson, 2002. Identifying reservoirs of infection: a conceptual and practical challenge. *Emerging infectious diseases* **8**:1468–73.
- Heffernan, J. M., R. J. Smith, and L. M. Wahl, 2005. Perspectives on the basic reproductive ratio. *Journal of the Royal Society, Interface / the Royal Society* **2**:281–93.
- Hethcote, H. W., 2000. The mathematics of infectious diseases. *SIAM Review* **42**:599–653.
- Hilker, F. M., M. Langlais, and H. Malchow, 2009. The Allee effect and infectious diseases: extinction, multistability, and the (dis-)appearance of oscillations. *The American Naturalist* **173**:72–88.
- Hogeweg, P., 1988. Cellular automata as a paradigm for ecological modeling. *Applied Mathematics and Computation* **27**:81–100.
- Holland, E. P., J. N. Aegerter, C. Dytham, and G. C. Smith, 2007. Landscape as a Model: The Importance of Geometry. *PLoS Computational Biology* **3**:e200.
- Holling, C. S., 1978. Adaptive environmental assessment and management. Wiley-Interscience.

- Ingerson, T. E. and R. L. Buvel, 1984. Structure in asynchronous cellular automata. *Physica D: Nonlinear Phenomena* **10**:59–68.
- Jeger, M. J. and F. van den Bosch, 1993. Threshold criteria for model plant disease epidemics. II. Persistence and endemicity. *Phytopathology* **84**:28–30.
- Jesse, M. and H. Heesterbeek, 2011. Divide and conquer? Persistence of infectious agents in spatial metapopulations of hosts. *Journal of Theoretical Biology* **275**:12–20.
- Jorgensen, S. E., 2008. Overview of the model types available for development of ecological models. *Ecological Modelling* **215**:3–9.
- Kamo, M. and M. Boots, 2004. The curse of the pharaoh in space: free-living infectious stages and the evolution of virulence in spatially explicit populations. *Journal of Theoretical Biology* **231**:435–441.
- Keeling, M. J. and C. A. Gilligan, 2000. Bubonic plague: a metapopulation model of a zoonosis. *Proceedings. Biological sciences / The Royal Society* **267**:2219–30.
- Keeling, M. J. and P. Rohani, 2007. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, first edition.
- Kermack, W. O. and A. G. McKendrick, 1927. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **115**:700–721.
- Kullback, S., 1987. The Kullback-Leibler distance. *The American Statistician* **41**:338–341.
- Kullback, S. and R. A. Leibler, 1951. On information and sufficiency. *Annals of Mathematical Statistics* **22**:79–86.
- Kurtz, T. G., 1976. Limit theorems and diffusion approximations for density dependent Markov chains. In R. J.-B. Wets, editor, *Stochastic Systems: Modeling, Identification and Optimization, I Mathematical Programming Studies, Volume 5*, volume 5 of *Mathematical Programming Studies*, pages 67–78. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Lafferty, K. D., 2009. The ecology of climate change and infectious diseases. *Ecology* **90**:888–900.
- Laird, R. A. and B. S. Schamp, 2008. Does local competition increase the coexistence of species in intransitive networks? *Ecology* **89**:237–247.
- Lankester, B. F., K. Hampson, T. Lembo, G. Palmer, and S. Cleaveland, 2014. Implementing Pasteur’s vision for rabies elimination. *Science* **345**:1562–1564.

- Lehoucq, R. B., D. C. Sorensen, and C. Yang, 1998. ARPACK User's Guide, Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. SIAM, Philadelphia, PA.
- Levin, S. A., 1974. Dispersion and Population Interactions. *The American Naturalist* **108**:207–228.
- Levins, R., 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the Entomological Society of America* **15**:237–240.
- Levins, R., 1970. Extinction. *Journal of the American Mathematical Society* **2**:77–107.
- Li, M. Y., J. R. Graef, L. Wang, and J. Karsai, 1999. Global dynamics of a SEIR model with varying total population size. *Mathematical Biosciences* **160**:191–213.
- Lion, S., 2010. Evolution of reproductive effort in viscous populations: the importance of population dynamics. *Journal of Evolutionary Biology* **23**:866–874.
- Lion, S. and M. van Baalen, 2007. From infanticide to parental care: why spatial structure can help adults be good parents. *The American Naturalist* **170**:E26–46.
- Lion, S. and M. van Baalen, 2008. Self-structuring in spatial evolutionary ecology. *Ecology Letters* **11**:277–295.
- Lloyd, A., 2001. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theoretical Population Biology* **60**:59–71.
- Lloyd-Smith, J. O., P. C. Cross, C. J. Briggs, M. Daugherty, W. M. Getz, J. Latta, M. S. Sanchez, A. B. Smith, and A. Swei, 2005. Should we expect population thresholds for wildlife disease? *Trends in Ecology & Evolution* **20**:511–9.
- Ludwig, D., 1999. Is It Meaningful to Estimate a Probability of Extinction? *Ecology* **80**:298–310.
- Lumer, E. D. and G. Nicolis, 1994. Synchronous versus asynchronous dynamics in spatially distributed systems. *Physica D: Nonlinear Phenomena* **71**:440–452.
- Mancy, R., P. Prosser, and S. Rogers, 2013. Discrete and continuous time simulations of spatial ecological processes predict different final population sizes and interspecific competition outcomes. *Ecological Modelling* **259**:50–61.
- Mangan, S. A., S. A. Schnitzer, E. A. Herre, K. M. L. Mack, M. C. Valencia, E. I. Sanchez, and J. D. Bever, 2010. Negative plant-soil feedback predicts tree-species relative abundance in a tropical forest. *Nature* **466**:752–5.

- Manzoni, L., 2012. Asynchronous cellular automata and dynamical properties. *Natural Computing* **11**:269–276.
- Matsuda, H., N. Ogita, A. Sasaki, and K. Sato, 1992. Statistical Mechanics of Population. *Progress of Theoretical Physics* **88**:1035–1049.
- McDonald-Madden, E., P. W. J. Baxter, R. A. Fuller, T. G. Martin, E. T. Game, J. Montambault, and H. P. Possingham, 2010. Monitoring does not always count. *Trends in Ecology & Evolution* **25**:547–550.
- McGeoch, C. C., 2007. Experimental algorithmics. *Communications of the ACM* **50**:27.
- McVinish, R. and P. K. Pollett, 2013a. Interaction between habitat quality and an Allee-like effect in metapopulations. *Ecological Modelling* **249**:84–89.
- McVinish, R. and P. K. Pollett, 2013b. The deterministic limit of a stochastic logistic model with individual variation. *Mathematical Biosciences* **241**:109–114.
- McVinish, R. and P. K. Pollett, 2013c. The limiting behaviour of a stochastic patch occupancy model. *Journal of mathematical biology* **67**:693–716.
- Meier, C. M., J. Starrfelt, and H. Kokko, 2011. Mate limitation causes sexes to coevolve towards more similar dispersal kernels. *Oikos* **120**:1459–1468.
- Méléard, S. and D. Villemonais, 2012. Quasi-stationary distributions and population processes. *Probability Surveys* **9**:340–410.
- Melesse, D. Y. and A. B. Gumel, 2010. Global asymptotic properties of an SEIRS model with multiple infectious. *Journal of Mathematical Analysis and Applications* **366**:202–217.
- Moilanen, A. and I. Hanski, 1995. Habitat Destruction and Coexistence of Competitors in a Spatially Realistic Metapopulation Model. *The Journal of Animal Ecology* **64**:141.
- Nasell, I., 1991. On the quasi-stationary distribution of the Ross malaria model. *Mathematical Biosciences* **107**:187–207.
- Nasell, I., 1995. The threshold concept in stochastic epidemic and endemic models. In D. Mollison, editor, *Epidemic models: their structure and relations to data*, pages 71–83. Cambridge University Press.
- Nasell, I., 1996. The Quasi-Stationary Distribution of the Closed Endemic SIS Model. *Advances in Applied Probability* **28**:895–932.

- Nasell, I., 1999. On the quasi-stationary distribution of the stochastic logistic epidemic. *Mathematical Biosciences* **156**:21–40.
- Nasell, I., 2001a. Extinction and Quasi-stationarity in the Verhulst Logistic Model. *Journal of Theoretical Biology* **211**:11–27.
- Nasell, I., 2001b. Extinction and quasi-stationarity in the Verhulst logistic model: with derivations of mathematical results (extended version). *Journal of Theoretical Biology* **211**:11–27.
- Nasell, I., 2005. A new look at the critical community size for childhood infections. *Theoretical Population Biology* **67**:203–16.
- National Institute of Health - Bioinformatics Definition Committee, 2000. Nih Working Definition of Bioinformatics and Computational Biology. Technical report.
- Odenbaugh, J., 2005. Idealized, Inaccurate but Successful: A Pragmatic Approach to Evaluating Models in Theoretical Ecology. *Biology & Philosophy* **20**:231–255.
- Onstad, D. W. and E. A. Kornkven, 1992. Persistence and endemicity of pathogens in plant populations over time and space. *Phytopathology* **82**:561–566.
- Orzack, S. H., 2012. The philosophy of modelling or does the philosophy of biology have any use? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **367**:170–80.
- Ovaskainen, O., 2003. Habitat destruction, habitat restoration and eigenvector-eigenvalue relations. *Mathematical Biosciences* **181**:165–76.
- Ovaskainen, O. and B. T. Grenfell, 2003. Mathematical tools for planning effective intervention scenarios for sexually transmitted diseases. *Sexually Transmitted Diseases* **30**:388–94.
- Ovaskainen, O. and I. Hanski, 2001. Spatially structured metapopulation models: global and local assessment of metapopulation capacity. *Theoretical Population Biology* **60**:281–302.
- Ovaskainen, O. and I. Hanski, 2003a. Extinction threshold in metapopulation models. *Annales Zoologici Fennici* **40**:81–97.
- Ovaskainen, O. and I. Hanski, 2003b. How much does an individual habitat fragment contribute to metapopulation dynamics and persistence? *Theoretical Population Biology* **64**:481–495.
- Ovaskainen, O. and I. Hanski, 2004. From individual behavior to metapopulation dynamics: unifying the patchy population and classic metapopulation models. *The American Naturalist* **164**:364–77.

- Ovaskainen, O. and B. Meerson, 2010. Stochastic models of population extinction. *Trends in Ecology & Evolution* **25**:643–652.
- Pallansch, M. A. and H. S. Sandhu, 2006. The eradication of polio—progress and challenges. *The New England journal of medicine* **355**:2508–11.
- Patz, J. A., D. Campbell-Lendrum, T. Holloway, and J. A. Foley, 2005. Impact of regional climate change on human health. *Nature* **438**:310–7.
- Perfecto, I. and J. Vandermeer, 2008. Spatial pattern and ecological process in the coffee agroforestry system. *Ecology* **89**:915–920.
- Pincock, C., 2012. Mathematical models of biological patterns: Lessons from Hamiltons selfish herd. *Biology & Philosophy* pages 481–496.
- Polack, F., 2014. Filling gaps in simulation of complex systems : the background and motivation for CoSMoS. *Natural Computing* .
- Pollett, P., 1999. Modelling quasi-stationary behaviour in metapopulations. *Mathematics and Computers in Simulation* **48**:393–405.
- Pollett, P. K., 1996. Modelling the long-term behaviour of evanescent ecological systems. *Ecological Modelling* **86**:135–139.
- Pollett, P. K., 2001. Quasi-stationarity in populations that are subject to large-scale mortality or emigration. *Environment International* **27**:231–6.
- Pollett, P. K., 2012. Quasi-stationary Distributions : A Bibliography. Technical Report July, University of Queensland.
- Pollett, P. K. and D. E. Stewart, 1994. An efficient procedure for computing quasi-stationary distributions of Markov chains with sparse transition structure. *Advances in Applied Probability* **26**:68–79.
- Post, W., D. DeAngelis, and C. Travis, 1983. Endemic disease in environments with spatially heterogeneous host populations. *Mathematical Biosciences* **63**:289–302.
- President’s Information Technology Advisory Committee, 2005. Computational Science: Ensuring America’s Competitiveness. Technical report.
- R Core Team, 2012. R: A Language and Environment for Statistical Computing.
- Ramaswamy, R., N. González-Segredo, and I. F. Sbalzarini, 2009. A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks. *The Journal of chemical physics* **130**:244104.

- Roberts, M. G., 2007. The pluses and minuses of R_0 . *Journal of the Royal Society, Interface / the Royal Society* **4**:949–61.
- Roxburgh, S. H., K. Shea, and J. B. Wilson, 2004. The intermediate disturbance hypothesis: patch dynamics and mechanisms of species coexistence. *Ecology* **85**:359–371.
- Ruxton, G. D. and L. A. Saravia, 1998. The need for biological realism in the updating of cellular automata models. *Ecological Modelling* **107**:105–112.
- Sanassy, D., P. Widera, and N. Krasnogor, 2014. Meta-stochastic Simulation of Biochemical Models for Systems & Synthetic Biology. *ACS Synthetic Biology* page 140822150814002.
- Schönfisch, B. and A. de Roos, 1999. Synchronous and asynchronous updating in cellular automata. *Biosystems* **51**:123–143.
- SIAM Working Group on CSE Education, 1998. Graduate Education for Computational Science and Engineering. Technical report.
- Slepoy, A., A. P. Thompson, and S. J. Plimpton, 2008. A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks. *Journal of Chemical Physics* **128**:1–8.
- Stankey, G. H., R. N. Clark, and B. T. Bormann, 2005. Adaptive Management of Natural Resources : Theory , Concepts , and Management Institutions. Technical report, United States Department of Agriculture.
- Steinsaltz, D. and S. N. Evans, 2004. Markov mortality models: implications of quasistationarity and varying initial distributions. *Theoretical Population Biology* **65**:319–37.
- Stepney, S. and P. S. Andrews, 2015. CoSMoS special issue editorial. *Natural Computing* .
- Stundzia, A. B. and C. J. Lumsden, 1996. Stochastic Simulation of Coupled Reaction - Diffusion Processes. *Journal of Computational Physics* **127**:196–207.
- Sun, S., C. Guo, and C. Li, 2012. Global Analysis of an SEIRS Model with Saturating Contact Rate. *Applied Mathematical Sciences* **6**:3991–4003.
- Tabares, P. C. C. and J. D. Ferreira, 2011. Weak Allee effect in a predator-prey system involving distributed delays. *Computational and Applied Mathematics* **30**:675–699.
- Takacs, P. and M. Ruse, 2011. The Current Status of the Philosophy of Biology. *Science & Education* **22**:5–48.
- Tedre, M., 2011. Computing as a science: A survey of competing viewpoints. *Minds and Machines* **21**:361–387.

- Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. N. Erasmus, M. F. De Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A. S. Van Jaarsveld, G. F. Midgley, L. Miles, M. a. Ortega-Huerta, a. T. Peterson, O. L. Phillips, and S. E. Williams, 2004. Extinction risk from climate change. *Nature* **427**:145–8.
- van den Driessche, P. and J. Watmough, 2002. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences* **180**:29–48.
- van Doorn, E. A. and P. K. Pollett, 2013. Quasi-stationary distributions for discrete-state models. *European Journal of Operational Research* **230**:1–14.
- Vandermeer, J., I. Perfecto, and S. M. Philpott, 2008. Clusters of ant colonies and robust criticality in a tropical agroecosystem. *Nature* **451**:457–9.
- Vandermeer, J. and S. Yitbarek, 2012. Self-organized spatial pattern determines biodiversity in spatial competition. *Journal of Theoretical Biology* **300**:48–56.
- Vere-Jones, D., 1969. Some limit theorems for evanescent processes. *Australian Journal of Statistics* **11**:67–78.
- Viana, M., R. Mancy, R. Biek, S. Cleaveland, P. C. Cross, J. O. Lloyd-Smith, and D. T. Haydon, 2014. Assembling evidence for identifying reservoirs of infection. *Trends in Ecology & Evolution* **29**:270–279.
- White, D. and A. R. Kiester, 2008. Topology matters: Network topology affects outcomes from community ecology neutral models. *Computers, Environment and Urban Systems* **32**:165–171.
- Wickham, H., 2009. ggplot2: elegant graphics for data analysis. Springer New York.
- Wilkinson, R. R. and K. J. Sharkey, 2013. An exact relationship between invasion probability and endemic prevalence for Markovian SIS dynamics on networks. *PLoS One* **8**:e69028.
- Wilson, G., D. a. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson, 2014. Best Practices for Scientific Computing. *PLoS Biology* **12**.
- Wilson, M. E., 2004. Travel and the emergence of infectious diseases. *Journal of agromedicine* **9**:161–77.
- Wright, S., 1931. Evolution in Mendelian Populations. *Genetics* **16**:97–159.
- Yaglom, A. M., 1947. Certain limit theorems of the theory of branching processes (in Russian). *Dokl. Acad. Nauk SSSR* **56**:795–798.

Zhou, S.-R. and G. Wang, 2004. Allee-like effects in metapopulation dynamics. *Mathematical Biosciences* **189**:103–13.